Scientific Reports

https://doi.org/10.1038/s41598-026-54121-y

Article in Press

Bootstrap-based K-means feature selection strategy for fuzzy regression functions

Received: 21 February 2026

Accepted: 18 May 2026

Published online: 21 May 2026

Cite this article as: Ucan A., Yildiz D. & Tak N. Bootstrap-based K-means feature selection strategy for fuzzy regression functions. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-54121-y

Aylin Ucan, Dogan Yildiz & Nihat Tak

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

ARTICLE IN PRESS

Bootstrap-Based K-Means Feature Selection Strategy for Fuzzy Regression Functions

Aylin Ucan[1,2], Dogan Yildiz[1], and Nihat Tak[2,*]

[1]Yildiz Technical University, Department of Statistics, Istanbul, 34220, Turkiye
[2]Marmara University, Department of Statistics, Istanbul, 34730, Turkiye
[*]nihat.tak@marmara.edu.tr ; nihattak@gmail.com

ABSTRACT

In recent years, the increasing complexity of datasets has highlighted the limitations of traditional feature selection methods and motivated the search for more stable and robust alternatives. This study proposes a bootstrap-stabilized feature selection method based on the k-means clustering algorithm, where features are repeatedly grouped, and stable representative features are identified across multiple resamples. In the proposed approach, each feature is assigned to its nearest cluster center, and the feature closest to that center is selected as the representative, reducing sensitivity to random initialization and improving generalizability. The selected features are subsequently integrated into a Fuzzy Regression Function (FRF) model, forming a hybrid framework that combines interpretable feature selection with flexible fuzzy regression modeling. The method is evaluated on ten real-world datasets and compared with widely used feature selection techniques, including Correlation based Feature Selection (CFS), Recursive Feature Elimination (RFE), Genetic Algorithm (GA) based Feature Selection, LASSO, and Ridge regression. In the proposed framework, the number of feature clusters was primarily determined using a grid-search strategy, while the Silhouette Index (SI) and Davies Bouldin Index (DBI) were additionally considered as alternative cluster validity criteria for k-means based feature selection. These findings demonstrate that the proposed bootstrap-stabilized clustering approach achieved either superior or competitive performance across the ten datasets considered, highlighting its effectiveness and suitability for complex data environments.

ARTICLE IN PRESS

1 Introduction

Many contemporary datasets contain large numbers of features and often display complex structures, making it difficult for regression models to perform reliably without some form of dimensionality reduction. Feature selection has therefore become an essential step in modern data analysis, helping to remove irrelevant or redundant features, reduce model complexity, and improve predictive accuracy [1]. By focusing on the most informative attributes, feature selection methods also mitigate overfitting and enhance interpretability, which is increasingly important in applied machine learning.

A wide range of feature selection strategies has been proposed in the literature. Comprehensive reviews in the literature [1, 2, 3] categorize these techniques into filter, wrapper, embedded, hybrid, and ensemble approaches. Filter methods select features based on statistical criteria and operate independently of the final predictive model [4]. Wrapper methods evaluate subsets of features using a trained model, often achieving good performance but at the cost of increased computation [5]. Embedded approaches, such as those used in LASSO [6] and Ridge regression [7], perform feature selection during model training and exploit algorithm specific regularization structures [8]. Hybrid methods combine complementary aspects of filters and wrappers to balance efficiency and accuracy [3], while ensemble based feature selection aims to aggregate multiple feature subsets to improve robustness [9]. Despite their widespread use, these methods may struggle when the feature space contains highly correlated features, leading to instability in the selected subsets or unnecessary redundancy.

As an alternative, clustering based feature selection has gained attention due to its simplicity and scalability. In particular, k-means clustering offers a straightforward way to group similar features and choose a representative from each cluster. However, classical k-means suffers from sensitivity to initialization and may produce inconsistent results across different runs, especially in high-dimensional or noisy settings. Bootstrap resampling provides a natural mechanism to improve stability by aggregating information across repeated partitions of the feature space. Although bootstrap techniques have been successfully applied in various machine learning contexts, their use within clustering driven feature selection has received limited attention.

ARTICLE IN PRESS

To address these issues, we propose a bootstrap-stabilized k-means feature selection algorithm that repeatedly clusters bootstrap-resampled datasets and identifies features that consistently appear nearest to their cluster centers. This approach reduces the randomness inherent in k-means and yields a more reliable and interpretable subset of representative features. The selected features are then incorporated into a Fuzzy Regression Function (FRF) model, originally introduced by [10], which provides a flexible framework for modeling uncertainty through membership based weighting. By employing bootstrap-based clustering in the FRF modelling structure, we aim to construct a feature selection and regression framework that is both interpretable and robust. The bootstrap mechanism stabilizes the k-means step, while the FRF model captures local variations in the relationship between predictors and the response feature. The resulting hybrid methodology reduces the dimensionality of the problem, lowers the computational cost of FRF parameter tuning, and improves overall predictive performance. In this study, the proposed method is evaluated on multiple real-world datasets and compared with widely used feature selection techniques to demonstrate its stability and effectiveness.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature. Section 3 presents the proposed methodology and its algorithmic framework. Section 4 reports the experimental results and comparative analyses. Section 5 discusses the main findings, and Section 6 concludes the study and outlines potential directions for future research.

2 Literature Survey

This section reviews several methodologies proposed in the literature that integrate feature selection and clustering strategies, particularly k-means based approaches, to improve clustering effectiveness.

Hancer et al. [11] presented a comprehensive survey on feature selection approaches for clustering, systematically reviewing existing unsupervised and evolutionary computation based methods while highlighting their strengths, limitations, and potential future research directions. Alimoussa et al. [12] proposed a clustering based sequential feature selection framework that removes irrelevant features, groups redundant ones through clustering, and sequentially selects representative features, demonstrating effective and efficient performance compared with benchmark methods. Xu et al. [13] proposed a multi-scale supervised clustering based feature selection algorithm (MCBFS) for genomic data analysis, demonstrating robust and effective performance in gene selection, classification, and biomarker identification compared with benchmark methods. Zhu et al. [14] proposed a joint spectral clustering framework with embedded feature selection and graph optimization, achieving superior clustering performance and improved robustness against redundant features and noise. Wang et al. [15] proposed a feature clustering assisted differential evolution based feature selection method that identifies multiple optimal feature subsets and improves classification accuracy while reducing feature subset size. Mei et al. [16] proposed a cooperative hybrid breeding swarm intelligence algorithm for feature selection, demonstrating superior convergence speed, classification accuracy, and feature subset stability on high dimensional benchmark datasets. Tripathi et al. [17] proposed a scalable distributed Laplacian Score framework for unsupervised feature selection, significantly improving computational efficiency and scalability while preserving the original method's accuracy and interpretability. Dwivedi et al. [18] presented a taxonomy of unsupervised feature selection methods, categorizing existing approaches and discussing their advantages, limitations, and practical challenges . Recent studies have also emphasized the importance of clustering quality evaluation through validity measures such as the Silhouette Index (SI) and Davies Bouldin Index (DBI). For example, Dwivedi et al. [19] proposed a scalable alignment free feature extraction framework for genomic data clustering by incorporating chemical property based entropy measures into clustering analysis . In another study, Dwivedi et al. [20] developed a multi-objective incremental clustering framework for dynamic data analysis by integrating inter cluster distance, intra cluster distance, and cluster density within a weighted optimization structure.

Table 1 summarizes representative k-means based feature selection approaches in the literature, highlighting their main methodological contributions.

ARTICLE IN PRESS

ARTICLE IN PRESS

Method	Methodological Approach
k-means clustering based feature selection	Uses k-means clustering to group redundant features and selects representative features from each cluster [21].
k-means UFS + ADMM	Formulates unsupervised feature selection directly through the k-means objective and solves it using an ADMM optimization framework [22].
Randomized feature selection for k-means	Introduces a randomized feature selection framework for k-means clustering with approximation guarantees [23].
Two-stage hybrid preprocessing for k-means	Combines cluster number estimation and feature selection as a preprocessing stage before applying k-means clustering [24].
ACO based clustering hybrid feature selection	Integrates ant colony optimization with k-means based quality measures to guide feature subset selection and improve exploration of the feature space [25].
ACO based hybrid feature selection	Uses ant colony optimization with k-means based evaluation criteria to assess feature quality and guide the search for representative feature subsets [26].

Table 1. Summary of representative k-means based feature selection approaches in the literature.

Based on the above literature, it can be observed that existing clustering based feature selection approaches mainly focus on optimization quality, scalability, or clustering performance. However, an important limitation remains in k-means based representative feature selection. Since k-means relies on random initialization of cluster centers, representative features selected in a single run may vary significantly across different executions, reducing the stability and robustness of the selected feature subset.

To address this limitation, this study proposes a bootstrap-stabilized k-means feature selection framework, where k-means clustering is repeatedly applied to bootstrap-resampled datasets and representative features are aggregated across multiple iterations. In this way, the proposed method aims to improve the stability of representative feature selection while preserving the simplicity and scalability advantages of k-means clustering. The selected features are subsequently integrated into the FRF modeling framework to improve predictive performance under uncertainty. FRF models have been successfully applied in a variety of regression problems [27, 28, 29, 30, 31, 32, 33, 34, 35], and have been further enhanced through regularization techniques such as Ridge based FRF [36, 37, 38] and Elastic Net FRF [39]. These studies demonstrate that FRF models can effectively capture nonlinear interactions and graded transitions in noisy environments.

ARTICLE IN PRESS

3 Proposed Method

The primary goal of the proposed method is to form a meaningful subset of features by grouping similar ones using the k-means clustering algorithm. Unlike conventional clustering based feature selection, the proposed approach employs a bootstrap-based repeated k-means strategy to alleviate sensitivity to random initialization. By selecting the most consistently observed representatives across multiple resampled runs, the method enhances robustness against randomness and improves the stability of the selected feature subset. In this way, the preprocessing phase of the prediction model is simplified, and its generalizability, interpretability, and predictive performance are improved. The k-means clustering algorithm was introduced by MacQueen (1967) and aims to partition data points into homogeneous subgroups based on their similarities. For k-means clustering analysis, the input matrix is constructed from the observation values in the dataset to be analyzed. Here, k denotes the number of clusters, and the algorithm proceeds through the following steps:

- **Step 1.** Data points are positioned in the feature space, and initial cluster centers are randomly selected.
- **Step 2.** Each data point is assigned to the cluster with the nearest center.
- **Step 3.** The center of each cluster is recalculated as the mean of the points assigned to that cluster.
- **Step 4.** Steps 2 and 3 are repeated until the cluster centers no longer change or a maximum number of iterations is reached.

ARTICLE IN PRESS

The distance measure is typically taken as Euclidean distance; however, alternative distance metrics that are frequently used in the literature may also be employed [40].

The proposed method is based on clustering the feature space in order to identify representative features from cluster of similar features. In the proposed framework, the training data matrix is first transposed so that each feature is represented as a vector in the observation space, and k-means clustering is applied to partition these feature vectors into homogeneous groups. A known limitation of the classical k-means algorithm is its sensitivity to the random initialization of cluster centers, which may lead to instability in the selected representative features across different runs. To reduce this sensitivity, the proposed approach incorporates a bootstrap-based resampling strategy [41], where bootstrap samples are repeatedly generated from the transposed training feature matrix and k-means clustering is independently applied at each bootstrap sampling. In each bootstrap sampling, the property with the smallest Euclidean distance from the center of each cluster is recorded as the representative property of that cluster. After completing all bootstrap samples, the final representative feature set is determined by selecting the most frequently occurring representative feature in each cluster across all bootstrap runs. This aggregation mechanism improves the stability, robustness, and representativeness of the selected feature subset while reducing the variance caused by random initialization.

Determining the optimal number of feature clusters in k-means remains a challenging problem, since there is no universally optimal cluster number for all datasets. Various cluster validity indices, such as the Elbow Method, Calinski Harabasz Index, Silhouette Index (SI), and Davies Bouldin Index (DBI), have been proposed in the literature to evaluate clustering quality [42, 43, 44]. Although the SI and DBI were employed in this study to determine the optimal number of clusters by evaluating cluster compactness and separation, these indices do not necessarily guarantee optimal predictive performance in subsequent regression tasks. Therefore, in this study, the primary methodology for determining the optimal number of feature clusters is based on a grid-search strategy over a predefined range of candidate cluster numbers. For each candidate value of k, the bootstrap-stabilized feature selection and FRF modelling procedure is executed, and the model with the minimum validation MAE is selected as the final model. Finally, using the optimal number of feature clusters k^*, the selected feature subset and the corresponding FRF model are applied to the test dataset to evaluate the final predictive performance of the proposed framework.

In line with these explanations, the flowchart (see Figure 1) and algorithmic steps (see Algorithm 1) of the proposed feature selection method are presented in Section 3.1.

3.1 Algorithm 1 : Bootstrapped k-means Feature Selection Method: FRF

Step 1: The data matrix X, which contains the independent features, and the response vector Y, representing the response feature, are partitioned into three subsets: training, validation, and testing.

$$X = [x_{ij}], \quad i = 1,2,\ldots,n;\ j = 1,2,\ldots,p \tag{1}$$

$$Y = [y_i], \quad i = 1,2,\ldots,n \tag{2}$$

$$X_{\text{train}} = [x_{ij}], \quad i = 1,2,\ldots,n_{\text{train}};\ j = 1,2,\ldots,p \tag{3}$$

$$Y_{\text{train}} = [y_i], \quad i = 1,2,\ldots,n_{\text{train}} \tag{4}$$

$$X_{\text{valid}} = [x_{ij}], \quad i = 1,2,\ldots,n_{\text{valid}};\ j = 1,2,\ldots,p \tag{5}$$

$$Y_{\text{valid}} = [y_i], \quad i = 1,2,\ldots,n_{\text{valid}} \tag{6}$$

$$X_{\text{test}} = [x_{ij}], \quad i = 1,2,\ldots,n_{\text{test}};\ j = 1,2,\ldots,p \tag{7}$$

ARTICLE IN PRESS

ARTICLE IN PRESS

$$Y_{\text{test}} = [y_i], \quad i = 1, 2, \ldots, n_{\text{test}} \tag{8}$$

Here, p denotes the number of features, and n indicates the total number of observations.

Step 2. The transpose of the training data matrix X_{train} is taken, denoted as X_{train}^T.

$$X_{\text{train}}^T = [x_{ji}], \quad j = 1, 2, \ldots, p;\ i = 1, 2, \ldots, n_{\text{train}} \tag{9}$$

Step 3: Feature selection based on the k-means clustering method.

The number of cluster (k) is primarily selected using a grid-search strategy over a predefined range of candidate values based on validation set performance. In addition, cluster validity indices such as the SI and DBI may also be considered as alternative criteria to guide the selection of an appropriate cluster number before proceeding with the bootstrap-based k-means feature selection process.

- **Step 3.1:** The maximum number of clusters ($k_{\max}$) and the maximum number of bootstrap iterations ($I_{\max}$) are defined.
- **Step 3.2:** The initial number of clusters is set to $k = 2$.
- **Step 3.3:** The bootstrap iteration index is initialized as $I = 1$.
  - **Step 3.3.1:** Using a bootstrap resampling approach from the transposed training data matrix X_{train}^T, a subset $X_{\text{train,selected}}^T$ is constructed by selecting a fixed number of observations of length n_{train}. This parameter represents the number of samples to be selected at each iteration and is defined as follows:

$$X_{\text{train,selected}}^T = [x_{ji}], \quad j = 1, 2, \ldots, p;\ i = 1, 2, \ldots, n_{\text{train}} \tag{10}$$

    This process is important to increase the generalizability of the model by performing analysis on representative samples of different subgroups in the data and to understand the behavior of cluster centers across different data groups.
  - **Step 3.3.2:** k initial cluster centers are determined from the matrix $X_{\text{train,selected}}^T$:

$$C^{(0)} = \{c_1^{(0)}, c_2^{(0)}, \ldots, c_k^{(0)}\}, \quad c_r^{(0)} \in X_{\text{train,selected}}^T,\ r = 1, 2, \ldots, k. \tag{11}$$

  - **Step 3.3.3:** For each feature vector x_j in the bootstrap sample $X_{\text{train,selected}}^T$, the distance to each of the k cluster centers is computed. The Euclidean distance between feature vector x_j and cluster center c_r is calculated as defined in Equation (12):

$$d(x_j, c_r) = \left\| x_j - c_r \right\|_2 = \sqrt{\sum_{i=1}^{n_{\text{train}}} (x_{ji} - c_{ri})^2}, \qquad j = 1, 2, \ldots, p,\ \ r = 1, 2, \ldots, k. \tag{12}$$

  - **Step 3.3.4:** Each feature vector x_j is assigned to the cluster whose center has the minimum Euclidean distance. The cluster index r^* denotes the nearest cluster center for feature vector x_j and is defined as in Equation (13):

$$r^* = \arg \min_{r=1,2,\ldots,k} d(x_j, c_r), \qquad j = 1, 2, \ldots, p. \tag{13}$$

    Thus, feature vector x_j is assigned to cluster C_{r^*}.

ARTICLE IN PRESS

/

ARTICLE IN PRESS

– **Step 3.3.5:** The cluster centers are updated by computing the mean of the feature vectors assigned to each cluster. The new center c_r of cluster r is calculated as shown in Equation (14).

$$c_r = \frac{1}{|C_r|} \sum_{j \in C_r} x_j, \qquad r = 1, 2, \ldots, k, \tag{14}$$

where C_r denotes the set of feature vectors assigned to cluster r, $x_j \in \mathbb{R}^{n_{\text{train}}}$ represents the j-th feature vector, and $c_r \in \mathbb{R}^{n_{\text{train}}}$ denotes the cluster center vector.

– **Step 3.3.6:** If the cluster centers do not change, proceed to Step 3.3.7; otherwise, return to Step 3.3.3.

– **Step 3.3.7:** In each cluster, the feature vector with the minimum distance to the cluster center is selected as the representative feature for the current bootstrap iteration. The selected representative feature indices are stored in the vector $m^{(I)}$:

$$m^{(I)} = \left[f_1^{(I)}, f_2^{(I)}, \ldots, f_k^{(I)} \right], \tag{15}$$

where $f_r^{(I)}$ denotes the index of the representative feature selected for cluster r in the I-th bootstrap iteration, corresponding to the feature vector with the minimum Euclidean distance to the cluster center c_r, for $r = 1, 2, \ldots, k$.

– **Step 3.3.8:** If $I = I_{\max}$, proceed to Step 3.4; otherwise, update the bootstrap iteration index as $I = I + 1$ and return to Step 3.3.1.

- **Step 3.4:** When $I = I_{\max}$, the matrix $M \in \mathbb{N}^{I_{\max} \times k}$ is constructed by stacking the representative feature indices obtained from all bootstrap iterations:

$$M = \begin{bmatrix} m^{(1)} \\ m^{(2)} \\ \vdots \\ m^{(I_{\max})} \end{bmatrix} = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1k} \\ f_{21} & f_{22} & \cdots & f_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ f_{I_{\max}1} & f_{I_{\max}2} & \cdots & f_{I_{\max}k} \end{bmatrix} \tag{16}$$

- **Step 3.5:** For each column of the matrix M, the most frequently selected feature index is determined. Thus, the final representative feature set is obtained as:

$$S = \{ f_1^*, f_2^*, \ldots, f_k^* \}, \tag{17}$$

where f_j^* denotes the final representative feature index for cluster j, determined as

$$f_j^* = \text{mode}(M_{1j}, M_{2j}, \ldots, M_{I_{\max} j}), \qquad j = 1, 2, \ldots, k. \tag{18}$$

- **Step 3.6:** Using the representative feature indices selected in Step 3.5 reduced subsets of the original matrices X_{train}, X_{valid}, and X_{test} are constructed as $X_{\text{train,reduced}}$, $X_{\text{valid,reduced}}$, and $X_{\text{test,reduced}}$, respectively.

- **Step 3.7:** The reduced dataset can be used in different regression models. In this study, the FRF model is employed, and the following steps describe the specific procedure of the FRF based regression framework. Here, c denotes the number of fuzzy clusters used in the FRF model and is distinct from k, which denotes the number of feature clusters in the k-means feature selection stage.

ARTICLE IN PRESS

ARTICLE IN PRESS

– **Step 3.7.1:** The optimal number of clusters (c) to be used in the FRF based approach is determined using the Fuzzy Silhouette Index (FSI) criterion.

– **Step 3.7.2:** The reduced training dataset $X_{\text{train,reduced}}$ is clustered using the fuzzy c-means clustering algorithm. For each cluster, the membership degree of each observation is computed based on the Euclidean distance defined in Equation (12), as shown in Equation (19):

$$\mu_{ir} = \left[\sum_{j=1}^{c} \left(\frac{d(x_{\text{train,reduced},r}, v_i)}{d(x_{\text{train,reduced},r}, v_j)} \right)^{\frac{2}{m-1}} \right]^{-1}, \quad i = 1,2,\ldots,c;\ r = 1,2,\ldots,n_{\text{train}} \tag{19}$$

Here, v_i denotes the center of cluster i, m is the fuzziness parameter, c is the number of clusters, and $d(\cdot)$ represents the Euclidean distance measure.

– **Step 3.7.3:** The membership degrees and the matrix of explanatory features are combined, and for the training dataset, the matrices corresponding to the i-th cluster are constructed as follows:

$$Y_{\text{train}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n_{\text{train}}} \end{bmatrix} \tag{20}$$

$$X^{(i)}_{\text{train,reduced}} = \begin{bmatrix} \mu_{i1} & x_{11} & \cdots & x_{p1} \\ \mu_{i2} & x_{12} & \cdots & x_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{in_{\text{train}}} & x_{1n_{\text{train}}} & \cdots & x_{pn_{\text{train}}} \end{bmatrix} \tag{21}$$

– **Step 3.7.4:** For each cluster, a regression model is constructed using Multiple Linear Regression (MLR). The coefficient vector $\hat{\beta}^{(i)}$ is predicted as shown in Equation (22):

$$\hat{\beta}^{(i)} = \arg\min_{\beta} \left\| Y^{(i)} - X^{(i)}_{\text{train,reduced}} \beta \right\|^2 \tag{22}$$

Once the coefficient vector $\hat{\beta}^{(i)}$ is calculated, the predictions on the validation set is computed as shown in Equation (23):

$$\hat{Y}^{(i)}_{\text{valid}} = X^{(i)}_{\text{valid,reduced}} \hat{\beta}^{(i)}, \qquad i = 1,2,\ldots,c \tag{23}$$

– **Step 3.7.5:** For the validation set, the final prediction values are calculated using the FRF based approach as shown in Equation (23). Specifically, the predictions obtained in Step 3.7.4 for each fuzzy cluster are weighted by the corresponding membership degrees (μ) obtained from Equation (24):

$$\hat{Y}_r = \frac{\sum_{i=1}^{c} \hat{Y}_{ir} \mu_{ir}}{\sum_{i=1}^{c} \mu_{ir}}, \qquad r = 1,2,\ldots,n_{valid}, \tag{24}$$

– **Step 3.7.6:** The results obtained from the FRF method are evaluated using the evaluation metric MAE and the corresponding metrics for the validation set are recorded.

ARTICLE IN PRESS

ARTICLE IN PRESS

- **Step 3.8:** If $k = k_{\max}$, proceed to Step 4; otherwise, set $k = k + 1$ and return to Step 3.3.

Step 4: Model selection and final test evaluation.

- **Step 4.1:** After repeating Steps 3.3–3.8 for all candidate cluster numbers $k = 2, 3, \ldots, k_{\max}$, the validation set prediction performance results are compared and the optimal number of clusters k^* is determined based on the minimum MAE value:

$$k^* = \arg \min_{k \in \{2,3,\ldots,k_{\max}\}} MAE_k \tag{25}$$

  where MAE_k denotes the validation MAE obtained for cluster number k.

- **Step 4.2:** Using the trained FRF model corresponding to the optimal cluster number k^*, predictions for the reduced test set as shown in Equatin (26):

$$\hat{Y}_{\text{test}}^{(i)} = X_{\text{test,reduced}}^{(i)} \hat{\beta}^{(i)}, \qquad i = 1, 2, \ldots, c \tag{26}$$

  where $\hat{\beta}^{(i)}$ denotes the coefficient vector estimated in Step 3.7.4.

- **Step 4.3:** The final test prediction values are calculated using the FRF based approach as shown in Equation (27).

$$\hat{Y}_r = \frac{\sum_{i=1}^{c} \hat{Y}_{ir} \mu_{ir}}{\sum_{i=1}^{c} \mu_{ir}}, \qquad r = 1, 2, \ldots, n_{\text{test}} \tag{27}$$

- **Step 4.4:** The predictive performance of the final selected model is evaluated on the test dataset using the evaluation metrics MAE.

ARTICLE IN PRESS

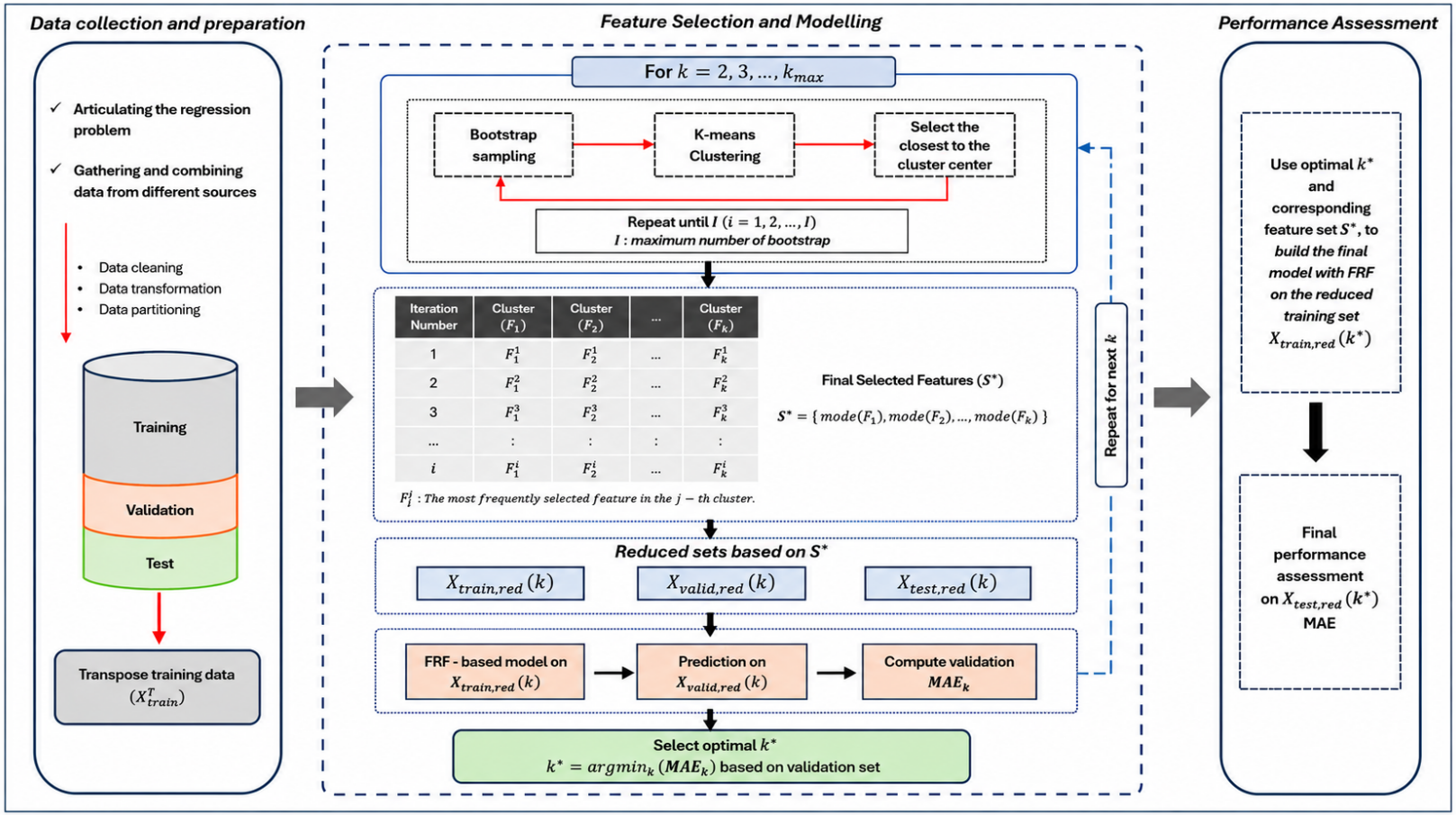

Figure 1. Flowchart of the proposed method.

ARTICLE IN PRESS

ARTICLE IN PRESS

4 Experimental Results

The performance of the proposed method is evaluated on ten real-world datasets with varying numbers of observations and features. Prior to analysis, all predictor features are standardized using Z-score normalization to ensure comparability across features and to improve the performance of clustering based feature selection. For each dataset, the observations are randomly partitioned into training (80%), validation (10%), and testing (10%) subsets. The training set is used for model prediction, the validation set is used for hyperparameter tuning and model selection, and the testing set is reserved for final performance evaluation. To ensure a fair comparison, the same data partitions are used for all competing methods within each dataset. All analyses are performed in the R programming environment.

In the proposed framework, the number of clusters (k) is treated as a tuning parameter and optimized iteratively. For datasets with fewer than 100 features, where p denotes the total number of features, the search interval is defined as $k = 2, 3, \ldots, p-1$, whereas for datasets with more than 100 features, the search interval is restricted to $k = 2, 3, \ldots, \lfloor p/2 \rfloor$ in order to reduce computational cost. For each candidate k, the proposed bootstrapped k-means feature selection procedure is applied to the training data. At each bootstrap sampling, k-means clustering is performed on the transposed training matrix, and representative features are identified as the features closest to the cluster centers. The final feature subset for each candidate k is determined by taking the mode of the representative feature indices obtained across bootstrap iterations for each cluster. Subsequently, the selected feature subset is used to construct the FRF model, and the predictive performance of each candidate k is evaluated on the validation dataset. The optimal number of clusters is then determined based on the minimum MAE obtained on the validation dataset, and the corresponding FRF model and selected feature subset are used for final evaluation on the test dataset.

For performance comparison, the proposed method is evaluated against both regression based and feature selection based benchmark methods, including Multiple Linear Regression (MLR), FRF, LASSO, Ridge Regression, Recursive Feature Elimination (RFE), Genetic Algorithm based Feature Selection (GA), and Correlation based Feature Selection (CFS). In the FRF model, the optimal number of fuzzy clusters (c) is determined using the FSI, while the fuzziness parameter is fixed at $m = 2$ throughout all analyses. The optimal cluster numbers are provided in Table 1 of the Supplementary Information. For LASSO and Ridge regression, the regularization parameter (λ) is determined by cross-validation using the `cv.glmnet` function from the `glmnet` package [45]. The optimal λ values obtained for each dataset are reported in Table 1 of the Supplementary Information. RFE and GA based feature selection procedures are implemented using the `caret` package [46]. For RFE, the `rfe` function is used, where the candidate subset sizes are defined over the range $1 : p$, with p representing the total number of independent features. The regression model is specified using the `"functions = lmFuncs"` function set. The optimal subset size selected for each dataset corresponds to the number of selected features reported in the feature reduction tables. For GA based feature selection, the `gafsControl` function is used with the underlying regression model specified as `"method = lm"`. The crossover probability is searched over the values $\{0.70, 0.80, 0.90\}$, the mutation probability over $\{0.01, 0.10\}$, and the elite proportion over the interval $[0, 1]$ using a grid-search strategy. The optimal parameter settings for each dataset are reported in Table 1 of the Supplementary Information. CFS is implemented using the `FSelector::cfs` function with its default settings [47].

A summary of the datasets, including the number of instances and features, is presented in Table 2. The first dataset, *Mtcars*, contains automobile data, including fuel consumption and design features for different vehicle models [48]. The second dataset, *BodyFat*, contains anthropometric measurements related to body composition, such as body fat percentage, waist circumference, and hip circumference obtained by dual-energy X-ray absorptiometry [49]. The third dataset, *College*, contains institutional and student level information for US universities, such as tuition, acceptance rates, graduation rates, and faculty characteristics [50]. The fourth dataset, *midwest*, provides demographic and socioeconomic indicators for various US states, including population density, poverty rates, and education levels [51]. The fifth dataset, *Automobile*, contains vehicle pricing and technical specifications [52]. The sixth dataset, *ethanol*, consists of near-infrared (*NIR*) spectral measurements collected in a chemical analysis study for ethanol prediction [53]. The seventh dataset, *gasoline*, contains *NIR* spectra of gasoline samples with known octane ratings [54]. The eighth dataset, *corn*, contains spectra of maize samples measured at multiple wavelengths for the prediction of chemical components [55]. The ninth dataset, *Wine*, contains physicochemical properties of white wines, including acidity, sugar content, sulfur dioxide levels, density, pH, and alcohol content, with the objective of predicting wine quality [56]. Finally, the tenth dataset, *Solubility*, available in the R environment, contains molecular descriptors of chemical compounds and is used to predict aqueous solubility levels, providing a benchmark for regression modeling in cheminformatics [57, 58].

To demonstrate the practical behavior of the proposed method, detailed feature selection results for the initial dataset

ARTICLE IN PRESS

are presented in Section 4.1. Results for the remaining datasets are summarized in terms of the number of features selected, and the prediction performance in the test dataset. In the results tables, BKMFS(I) represents the bootstrap version of the proposed method, where I is the number of bootstrap samples and KMFS represents the proposed feature selection method without bootstrap sampling. To examine the impact of the bootstrap size effect on performance stability, the proposed method is evaluated using bootstrapping sizes of 10, 25, 50, 100, 250, 500, and 1000. In the comparative result tables, only the best performing configuration of the proposed method for each dataset is included, and the MAE values corresponding to the remaining bootstrap numbers are presented in Table 23, while the numbers of selected features corresponding to these methods are reported in Table 2 of the Supplementary Information.

No	Dataset	Instances	Features
1	Mtcars	32	10
2	BodyFat	249	13
3	College	777	17
4	Midwest	437	24
5	Automobile	205	24
6	Ethanol	166	235
7	Gasoline	60	401
8	Corn	80	700
9	Wine	4898	11
10	Solubility	1267	228

Table 2. Summary of real-world datasets.

The methods are primarily evaluated using the MAE, since MAE provides an intuitive and scale dependent measure of prediction accuracy that is easy to interpret and less sensitive to issues such as division by small values, which can affect MAPE. In addition to MAE, RMSE and MAPE are also reported as complementary performance measures. The formulations of MAE, RMSE, and MAPE are provided in Equations 28, 29 and 30, respectively.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{28}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{29}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{30}$$

where y_i represents the actual values, $\hat{y}_i$ represents the predicted values, and n denotes the number of observations.

4.1 Mtcars dataset

The mtcars dataset is used in the first evaluation. This dataset consists of 32 instances and 10 independent features ("cyl", "disp", "hp", "drat", "wt", "qsec", "vs", "am", "gear", "carb"). In this dataset, the objective is to predict the miles per gallon (*mpg*) target feature. Table 3 presents the number of selected features and the corresponding feature names obtained by each feature selection method for the mtcars dataset.

KMFS represents the proposed feature selection procedure without bootstrap sampling. Figure 2, shows the representative features selected from each cluster in KMFS (for the case of k=6) and provides a subset of the baseline features for comparison with the bootstrap sampling. Figure 3, presents the representative features identified across 10 bootstrap samples for the case of $k = 6$, illustrating how the proposed method captures feature selection variability through repeated clustering. Figure 4, summarizes the mode of each feature being selected as a representative across all bootstraps. The mode selected feature within each cluster is then determined as the final representative, completing the feature selection process.

ARTICLE IN PRESS

Table 4, reports the prediction performance of the methods on the test dataset. Only the best-performing configuration of the proposed method is presented in the main result tables, whereas the remaining MAE values are reported in Table 23. According to the results, the lowest MAE values are obtained by BKMFS(10) with an MAE value of 2.1407. Among the benchmark methods, LASSO achieves the second-best MAE performance, whereas the proposed method provides an improvement of approximately 7.14% in MAE. Based on the RMSE and MAPE results, the proposed method achieves the best overall performance among the compared methods. Although no statistically significant differences are observed among the competing methods according to Supplementary Information Table 5, the lower prediction error obtained by the proposed method indicates a minor improvement in predictive performance.

Figure 5-(a), presents the sensitivity analysis of the proposed method for the mtcars dataset under different bootstrap numbers. The variation in MAE values across different bootstrap numbers indicates that the feature selection process is sensitive to the number of bootstrap, as different bootstrap numbers may lead to different feature subsets and consequently affect predictive performance. The use of bootstrap sampling contributed to an approximately 10.03% improvement in MAE for the mtcars dataset.

Method	Number of available features	Number of selected features	Selected features
MLR	10	–	–
FRF	10	–	–
RIDGE	10	–	–
LASSO	10	6	am, carb, drat, hp, qsec, wt
RFE	10	7	am, carb, cyl, drat, gear, qsec, wt
CFS	10	4	am, drat, qsec, vs
GA	10	4	am, carb, drat, hp
SI	10	3	disp, drat, carb
DBI	10	9	am, carb, drat, disp, qsec, gear, cyl, hp , wt
KMFS	10	6	am, carb, disp, drat, qsec, wt
BKMFS(10)	10	6	am, carb, cyl, drat, qsec, wt

Table 3. Feature reduction results of all methods for the *mtcars* dataset.

Method	MAE	RMSE	MAPE
MLR	4.1806	4.9327	0.1970
FRF	4.1375	5.6493	0.1536
RIDGE	2.4114	2.5397	0.1085
LASSO	2.3053	2.5494	0.1133
RFE	2.4716	2.7620	0.1223
CFS	2.4978	3.1405	0.1146
GA	2.5872	3.4120	0.1320
SI	2.9365	3.3350	0.1456
DBI	3.8363	4.4432	0.1764
KMFS	2.3794	2.5257	0.1120
BKMFS(10)	**2.1407**	**2.2788**	**0.1032**

* The bold values indicate the best result.

Table 4. Prediction performance comparison on the *mtcars* test set based on MAE, RMSE, and MAPE.

ARTICLE IN PRESS

ARTICLE IN PRESS

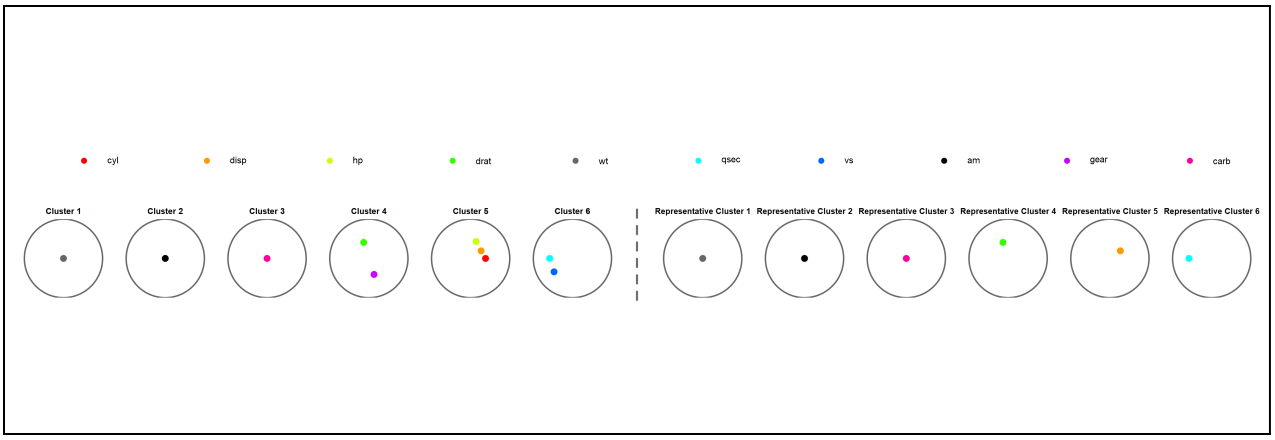

Figure 2. Cluster representatives for the *mtcars* dataset (non-bootstrap).

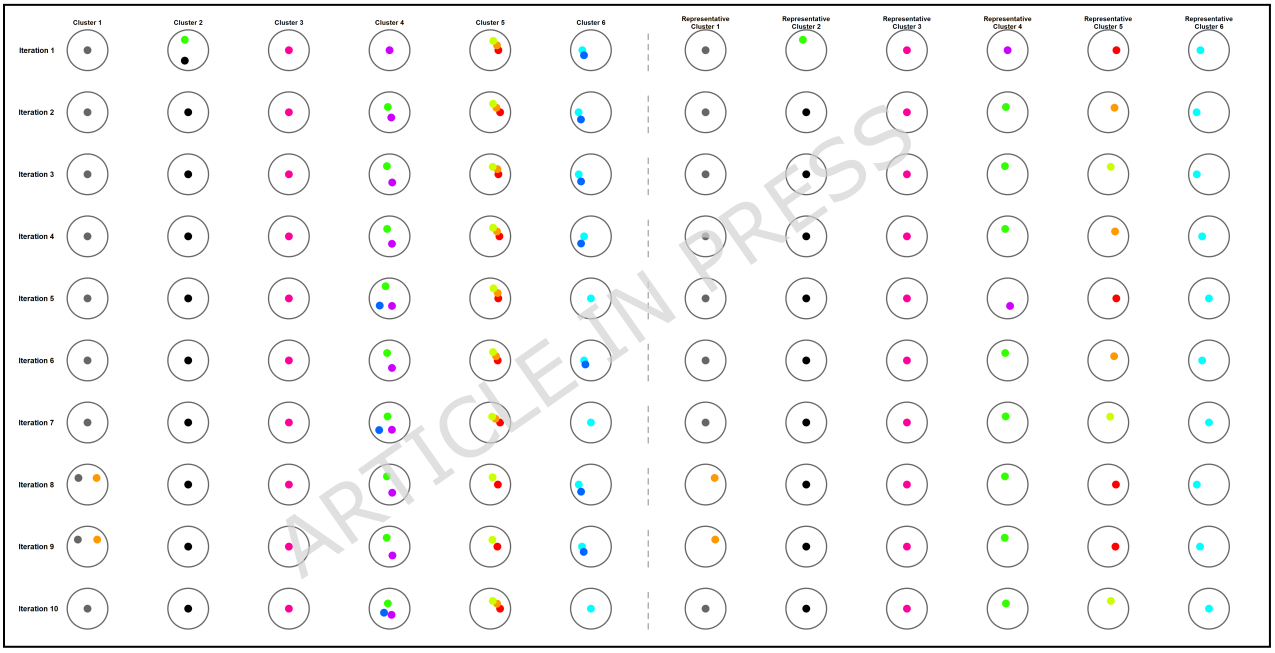

Figure 3. Cluster representatives across 10 iterations for the *mtcars* dataset.

ARTICLE IN PRESS

/

ARTICLE IN PRESS

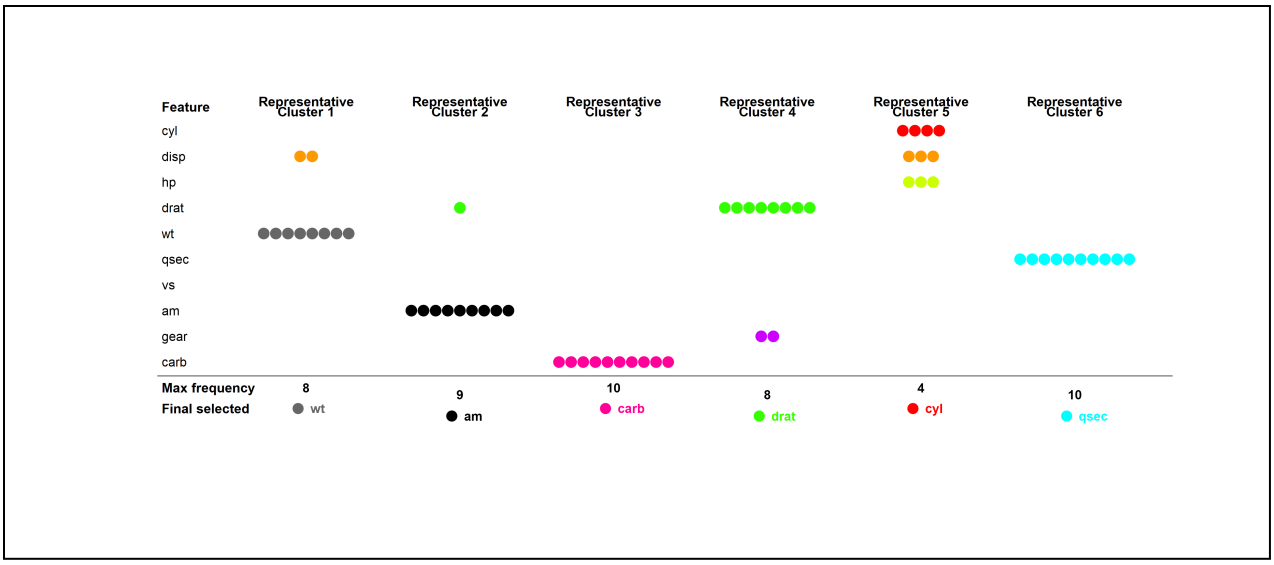

Figure 4. Cluster representative frequencies and final selected features for the *mtcars* dataset.

4.2 BodyFat dataset

To evaluate the performance of the proposed method, the BodyFat dataset is analyzed. The main goal of this dataset is to predict the *bodyfat* value. Table 5 presents the number of selected features and the corresponding feature names obtained by each feature selection method for the BodyFat dataset. Table 6 presents the prediction performance of the competing methods on the test dataset. Only the best-performing configuration of the proposed method is presented in the main result tables, whereas the remaining MAE values are reported in Table 23. According to the results, the lowest MAE value is obtained by BKMFS(500) with an MAE of 4.2002. Among the benchmark methods, the FRF model achieves a highly competitive performance with an MAE value of 4.2626; however, the proposed BKMFS(500) method performed better by reducing the MAE by approximately 1.46%. In terms of RMSE, BKMFS(10) achieves the best performance, further supporting the effectiveness of the proposed method in predictive modeling. Although DBI based k-means achieves the lowest MAPE value, the proposed BKMFS(500) method produces a highly competitive MAPE result with fewer selected features. In addition, no statistically significant differences are identified among the competing methods according to Supplementary Information Table 5, while the lower prediction error achieves by the proposed method suggests a minor improvement in predictive performance.

Figure 5-(b), presents the sensitivity analysis of the proposed method for the BodyFat dataset under different bootstrap numbers. The variation in MAE values across different bootstrap numbers indicates that the predictive performance is influenced by the number of bootstrap. Compared to the mtcars dataset, the results suggest that a larger number of bootstrap may be required to achieve the best predictive performance for the BodyFat dataset. The integration of bootstrap sampling yielded an approximately 7.89% reduction in MAE.

ARTICLE IN PRESS

ARTICLE IN PRESS

Method	Number of available features	Number of selected features	Selected features
MLR	13	–	–
FRF	13	–	–
RIDGE	13	–	–
LASSO	13	12	Age, Weight, Height, Neck, Chest, Abdomen, Hip, Thigh, Ankle, Biceps, Forearm, Wrist
RFE	13	13	Wrist, Abdomen, Neck, Forearm, Thigh, Hip, Biceps, Chest, Ankle, Knee, Height, Age, Weight
CFS	13	1	Abdomen
GA	13	7	Age, Weight, Height, Abdomen, Hip, Ankle, Wrist
SI	13	2	Chest, Weight
DBI	13	12	Chest, Thigh, Wrist, Ankle, Neck, Abdomen, Age, Weight, Height, Biceps, Knee, Forearm
KMFS	13	11	Chest, Thigh, Wrist, Ankle, Neck, Abdomen, Age, Hip, Height, Biceps, Knee
BKMFS(500)	13	8	Neck, Ankle, Wrist, Height, Forearm, Abdomen, Age, Hip

Table 5. Feature reduction results for all methods on the *BodyFat* dataset.

Method	MAE	RMSE	MAPE
MLR	4.3258	5.3635	0.2946
FRF	4.2626	5.2828	0.2899
RIDGE	4.2718	5.5997	0.3016
LASSO	4.3966	5.5943	0.3069
RFE	4.3258	5.3635	0.2946
CFS	5.0353	6.8597	0.3454
GA	4.4478	5.6386	0.3207
SI	4.7638	6.1089	0.3532
DBI	4.2522	5.3516	**0.2873**
KMFS	4.5601	5.7488	0.3103
BKMFS(500)	**4.2002**	**5.1794**	0.2984

* The bold values indicate the best result.

Table 6. Prediction performance comparison on the *BodyFat* test set based on MAE, RMSE, and MAPE.

ARTICLE IN PRESS

4.3 College dataset

To assess the effectiveness of the proposed method, the College dataset is analyzed, with the primary objective of predicting the *Grad.Rate* feature. Table 7 presents the number of selected features and the corresponding feature names obtained by each feature selection method for the College dataset, while Table 8 presents the prediction performance of the competing methods on the test dataset. Only the best-performing configuration of the proposed method is presented in the main result tables, whereas the remaining MAE values are reported in Table 23. According to the results, the lowest MAE value is achieved by the FRF method, with an MAE of 8.9462. BKMFS(50) and DBI based k-means achieves highly competitive performance with an MAE of 8.9486, ranking second overall.

It is noteworthy that RFE, GA, and the proposed method retain relatively large feature subsets compared to CFS. This behavior suggests that a substantial proportion of the original predictors contributes meaningful information to the response feature, and that excessive feature elimination may result in information loss. The relatively weaker performance of CFS, which retains only four features, further supports this interpretation. Moreover, although BKMFS(50) retain the same number of features as KMFS, the lower MAE value indicates that bootstrap stabilization contributed to selecting a more effective representative feature subset.

In terms of RMSE, the results also support the FRF method, while the lowest MAPE value is achieved by the MLR and GA models. However, as shown in Table 8, both MLR and GA retain the full feature set, indicating that their competitive predictive performance is achieved without any feature reduction.

Figure 5-(c), presents the sensitivity analysis of the proposed method for the College dataset under different bootstrap numbers. The results show that predictive performance improves progressively up to 50 bootstrap samples, suggesting that repeated bootstrap sampling enhances the effectiveness of feature selection in the early stages. However, a noticeable increase in MAE is observed at 100 bootstrap, indicating that additional bootstrap repetitions may introduce alternative representative feature subsets that are less effective for prediction. The incorporation of bootstrap sampling

ARTICLE IN PRESS

leads to an approximately 3.29% improvement in MAE compared with KMFS, indicating its positive contribution to feature subset quality and predictive performance.

Method	Number of available features	Number of selected features	Selected features
MLR	17	–	–
FRF	17	–	–
RIDGE	17	–	–
LASSO	17	12	Private, Apps, Top10perc, Top25perc, P.Undergrad, Outstate, Room.Board, Personal, PhD
RFE	17	14	3 Private, perc.alumni, S.F.Ratio, Top25perc, PhD, Terminal, Top10perc, Enroll, Room.Board ,Personal,P.Undergrad, Apps, Books, Accept
CFS	17	4	Top25perc, Outstate, Room.Board, perc.alumni
GA	17	17	Private, Apps, Accept, Enroll, Top10perc, Top25perc, F.Undergrad, P.Undergrad, Outstate, Room.Board, Books, Personal, PhD, Terminal, S.F.Ratio, perc.alumni, Expend
SI	17	8	Top25perc, Outstate, Private, Accept, P.Undergrad, Personal, Room.Board, S.F.Ratio
DBI	17	16	Top25perc, Outstate, perc.alumni, Accept, Enroll, Personal, Room.Board, S.F.Ratio, F.Undergrad, Private, Apps, PhD, P.Undergrad, Top10perc, Books, Terminal
KMFS	17	16	Top25perc, Expend, perc.alumni, Accept, Enroll, Personal, Room.Board, S.F.Ratio, F.Undergrad, Private, Apps, PhD, P.Undergrad, Top10perc, Books, Terminal
BKMFS(50)	17	16	Top25perc, Outstate, perc.alumni, Accept, Enroll, Personal, Room.Board, S.F.Ratio, F.Undergrad, Private, Apps, PhD, P.Undergrad, Top10perc, Books, Terminal

Table 7. Feature reduction results for all methods on the *College* dataset.

Method	MAE	RMSE	MAPE
MLR	8.9662	11.4182	**0.1372**
FRF	**8.9462**	**11.3625**	0.1374
RIDGE	9.1504	11.5769	0.1417
LASSO	9.1294	11.5570	0.1410
RFE	9.2735	11.6906	0.1432
CFS	9.7639	12.3370	0.1562
GA	8.9662	11.4182	**0.1372**
SI	9.2251	11.9410	0.1417
DBI	8.9486	11.4403	0.1379
KMFS	9.2532	11.6600	0.1426
BKMFS(50)	8.9486	11.4403	0.1379

* The bold values indicate the best result.

Table 8. Prediction performance comparison on the *College* test set based on MAE, RMSE, and MAPE.

4.4 Midwest dataset

For the Midwest dataset, the primary objective is to predict the *poverty* feature. Table 9 presents the number of selected features and the corresponding feature names obtained by each feature selection method, while Table 10 presents the prediction performance of the competing methods on the test dataset. The main result tables report only the best-performing configuration of the proposed method, while the remaining MAE results are given in Table 23. According to the results, the lowest MAE value is obtained by BKMFS(250), with an MAE of 0.1135, indicating the best predictive performance among all competing methods. Among the benchmark methods, FRF achieves the best performance with an MAE of 0.1199, whereas BKMFS(250) reduced the MAE by approximately 5.34%. In terms of RMSE and MAPE, BKMFS(250) also achieves the best results, further supporting the predictive effectiveness of the proposed method.

Figure 5-(d), presents the sensitivity analysis of the proposed method for the Midwest dataset under different bootstrap numbers. The results show that predictive performance stabilizes at 250 bootstrap samples and beyond,

ARTICLE IN PRESS

indicating that the feature selection process becomes more stable as the number of bootstrap increases. The incorporation of bootstrap sampling results in an approximately 5.35% improvement in MAE, highlighting its positive contribution to feature subset quality and predictive performance.

Method	Number of available features	Number of selected features	Selected features
MLR	24	–	–
FRF	24	–	–
RIDGE	24	–	–
LASSO	24	15	popdensity, popwhite, popasian, percblack, percamerindan, percasian, percother, popadults, perchsd, percprof, percpovertyknown, percchildbelowpovert, percadultpoverty, percelderlypoverty, category
RFE	24	16	percadultpoverty, area, percchildbelowpovert, percelderlypoverty, percamerindan, percwhite, percprof, percblack, percasian, perchsd, inmetro, category, percpovertyknown, percollege, popamerindian, popadults
CFS	24	2	percchildbelowpovert, percadultpoverty
GA	24	16	poptotal, popdensity, popblack, popamerindian, percwhite, percamerindan, popadults, perchsd, percprof, poppovertyknown, percpovertyknown, percchildbelowpovert, percadultpoverty, percelderlypoverty, inmetro, category
SI	24	6	percollege, popother, category, percchildbelowpovert, percwhite, poptotal
DBI	24	22	percasian, popasian, percblack, percelderlypoverty, percwhite, poptotal, percamerindan, area, popwhite, percchildbelowpovert, popdensity, category, popother, poppovertyknown, percother, inmetro, percadultpoverty, popadults, perchsd, popblack, popamerindian, percpovertyknown
KMFS	24	23	percasian, popasian, percblack, percelderlypoverty, percwhite, poptotal, percamerindan, area, popwhite, percchildbelowpovert, popdensity, category, popother, poppovertyknown, percother, inmetro, percadultpoverty, popadults, perchsd, popblack, popamerindian, percpovertyknown, percollege
BKMFS(250)	24	20	percasian, popasian, percblack, percelderlypoverty, percwhite, popamerindian, percamerindan, area, popwhite, percchildbelowpovert, popdensity, category, popother, poppovertyknown, percother, inmetro, percadultpoverty, popadults, perchsd, popblack

Table 9. Feature reduction results for all methods on the *Midwest* dataset.

ARTICLE IN PRESS

Method	MAE	RMSE	MAPE
MLR	0.1217	0.1671	0.0091
FRF	0.1199	0.1675	0.0089
RIDGE	0.2455	0.3237	0.0200
LASSO	0.1275	0.1786	0.0095
RFE	0.1291	0.1767	0.0097
CFS	0.4221	0.6057	0.0294
GA	0.1233	0.1674	0.0092
SI	0.8804	1.2135	0.0624
DBI	0.1155	0.1682	0.0087
KMFS	0.1198	0.1706	0.0090
BKMFS(250)	**0.1135**	**0.1647**	**0.0085**

* The bold values indicate the best result.

Table 10. Prediction performance comparison on the *Midwest* test set based on MAE, RMSE, and MAPE.

4.5 Automobile dataset

To investigate the predictive performance of the proposed method, the Automobile dataset is analyzed, with the primary objective of predicting the *price* feature. Table 11 presents the number of selected features and the corresponding feature names obtained by each feature selection method, while Table 12 presents the prediction performance of the competing methods on the test dataset. The main result tables report only the best-performing configuration of the proposed method, while the remaining MAE results are given in Table 23. According to the results, the lowest MAE value is achieved by BKMFS(10), with an MAE of 2365.4656, indicating the best predictive performance among all competing methods. Among the benchmark methods, LASSO achieves the best performance with an MAE of 2495.5840, whereas

ARTICLE IN PRESS

BKMFS(10) achieved an approximately 5.21% reduction in MAE. In addition, the DBI based k-means also achieves competitive performance with an MAE of 2398.1667. The RMSE results further support this finding, as BKMFS(10) also produces the lowest RMSE value among the competing methods. Although no statistically significant differences are identified among the competing methods according to Supplementary Information Table 5, the lower prediction error achieves by the proposed method suggests a slight improvement in predictive performance.

Figure 5-(e), presents the sensitivity analysis of the proposed method for the Automobile dataset under different bootstrap numbers. The results show that predictive performance improves substantially at lower bootstrap levels and becomes stable after 250 iterations. The incorporation of bootstrap sampling results in an approximately 12.19% improvement in MAE, highlighting its contribution to feature subset quality and predictive performance.

Method	Number of available features	Number of selected features	Selected features
MLR	24	–	–
FRF	24	–	–
RIDGE	24	–	–
LASSO	24	19	symboling, make, aspiration, num-of-doors, body-style, drive-wheels, engine-location, wheel-base, length, height, curb-weight, engine-type, engine-size, fuel-system, bore, stroke, compression-ratio, peak-rpm,highway-mpg
RFE	24	18	engine.location, fuel.type, bore, stroke, num.of.cylinders, aspiration, compression.ratio, drive.wheels, engine.type, body.style, num.of.doors, height, fuel.system, make, symboling, width, engine.size, highway.mpg
CFS	24	2	curb.weight, engine.size
GA	24	10	symboling, make, engine.location, wheel.base, length, curb.weight, engine.type, engine.size, fuel.system, city.mpg
SI	24	6	engine.size, city.mpg, aspiration, height, peak.rpm, length
DBI	24	23	horsepower, compression.ratio, bore, body.style, fuel.type, wheel.base, fuel.system, num.of.cylinders, engine.location, drive.wheels, engine.size, aspiration, symboling, engine.type, city.mpg, make, width, highway.mpg, num.of.doors, peak.rpm, stroke, length, curb.weight
KMFS	24	23	horsepower, compression.ratio, bore, height, fuel.type, wheel.base, fuel.system, num.of.cylinders, engine.location, drive.wheels, engine.size, aspiration, symboling, engine.type, city.mpg, make, width, highway.mpg, body.style, peak.rpm, stroke, length, curb.weight
BKMFS(10)	24	17	horsepower, compression.ratio, bore, body.style, fuel.type, wheel.base, fuel.system, num.of.cylinders, engine.location, drive.wheels, stroke, aspiration, symboling, engine.type, city.mpg, make, width

Table 11. Feature reduction results for all methods on the *Automobile* dataset.

ARTICLE IN PRESS

Method	MAE	RMSE	MAPE
MLR	2593.0360	3791.3160	0.2300
FRF	2606.9500	3805.8030	0.2306
RIDGE	2496.2150	3612.7510	0.2177
LASSO	2495.5840	3642.6600	0.2167
RFE	2614.1300	3780.2880	0.2411
CFS	2783.4860	4449.1470	0.2057
GA	2501.1810	3735.0130	0.2112
SI	2549.4790	3938.3280	**0.2025**
DBI	2398.1667	3573.6270	0.2118
KMFS	2693.8161	3898.4682	0.2384
BKMFS(10)	**2365.4656**	**3409.8820**	0.2090

* The bold values indicate the best result.

Table 12. Prediction performance comparison on the *Automobile* test set based on MAE, RMSE, and MAPE.

4.6 Ethanol dataset

For the Ethanol dataset, the primary objective is to predict the *ethanol concentration* feature. Table 13 presents the number of selected features and the corresponding feature names obtained by each feature selection method, while Table 14 presents the prediction performance of the competing methods on the test dataset. The main result tables

ARTICLE IN PRESS

report only the best-performing configuration of the proposed method, while the remaining MAE results are given in Table 23. According to the results, the lowest MAE value is achieved by BKMFS(1000), with an MAE of 0.9317, indicating the best predictive performance among all competing methods. Among the methods, LASSO achieves the second-best MAE performance with an MAE of 1.0872, whereas BKMFS(1000) reduced the MAE by approximately 14.30%. In terms of RMSE and MAPE, the proposed method also achieves the best predictive performance, further supporting its effectiveness for high-dimensional regression problems.

Figure 5-(f), presents the sensitivity analysis of the proposed method for the Ethanol dataset under different bootstrap numbers. The results show that the predictive performance varies substantially across bootstrap settings, with the best performance achieves at 1000 iterations. This behavior suggests that for high dimensional datasets, a larger number of bootstrap repetitions may be required to achieve more stable feature selection and improved predictive performance. Overall, the incorporation of bootstrap sampling leads to an approximately 36.05% improvement in MAE, demonstrating its contribution to more effective feature selection and improved predictive performance.

Method	**Number of available features**	**Number of selected features**	**Selected features**
MLR	235	–	–
FRF	235	–	–
RIDGE	235	–	–
LASSO	235	14	X1190.0, X1240.0, X1270.0, X1430.0, X1435.0, X1480.0, X1600.0, X1665.0, X1670.0, X1695.0, X1720.0, X2205.0, X2210.0,X2230.0
RFE	235	77	See Supplementary Information Table 8
CFS	235	2	X1670.0, X1675.0
GA	235	35	See Supplementary Information Table 8
SI	235	2	X1150.0, X1365.0
DBI	235	113	See Supplementary Information Table 8
KMFS	235	59	See Supplementary Information Table 8
BKMFS(1000)	235	63	See Supplementary Information Table 8

Table 13. Feature reduction results for all methods on the *Ethanol* dataset.

Method	**MAE**	**RMSE**	**MAPE**
MLR	25.7546	36.8162	0.4005
FRF	24.2294	36.0756	0.3798
RIDGE	4.6626	5.3338	0.0901
LASSO	1.0872	1.3847	0.0191
RFE	2.6468	3.5277	0.0397
CFS	2.3074	2.9270	0.0467
GA	1.2126	1.5148	0.0184
SI	6.4451	7.8515	0.1296
DBI	1.9582	2.8727	0.0309
KMFS	1.4570	1.7536	0.0280
BKMFS(1000)	**0.9317**	**1.1956**	**0.0160**

* The bold values indicate the best result.

Table 14. Prediction performance comparison on the *Ethanol* test set based on MAE, RMSE, and MAPE.

4.7 Gasoline dataset

To assess the effectiveness of the proposed method, the Gasoline dataset is examined. The main objective of this dataset is to predict the *octane number* feature. Table 15 presents the number of selected features and the corresponding feature names obtained by each feature selection method. Table 16 presents the prediction performance of the competing methods on the test dataset. The main result tables report only the best-performing configuration of the proposed method, while the remaining MAE results are given in Table 23. According to the results, the lowest MAE value is achieved by the LASSO method with an MAE of 0.2083. The RMSE and MAPE results also support this finding, confirming LASSO as the best-performing method for the Gasoline dataset.

Among the proposed method variants, BKMFS(50) achieves the best MAE performance with an MAE of 0.2142, ranking second overall among all competing methods. It is noteworthy that LASSO achieves its performance by

ARTICLE IN PRESS

selecting 17 features, whereas BKMFS(50) achives a highly competitive predictive performance using only 13 features, indicating a more compact feature subset.

Figure 5-(g), presents the sensitivity analysis of the proposed method for the Gasoline dataset under different bootstrap numbers. The results indicate that predictive performance improves up to 50 bootstrap, after which performance gradually deteriorates, suggesting that a moderate number of bootstrap repetitions provides the most effective feature selection behavior for this dataset. The incorporation of bootstrap sampling results in an approximately 30.36% improvement in MAE, highlighting its contribution to feature subset quality and predictive performance.

Method	Number of available features	Number of selected features	Selected features
MLR	401	–	–
FRF	401	–	–
RIDGE	401	–	–
LASSO	401	17	x.914.nm, x.1208.nm, x.1210.nm, x.1212.nm, x.1214.nm, x.1218.nm, x.1224.nm, x.1228.nm, x.1362.nm, x.1364.nm, x.1366.nm, x.1534.nm, x.1636.nm, x.1682.nm, x.1690.nm, x.1692.nm, x.1694.nm
RFE	401	13	x.948.nm, x.956.nm, x.958.nm, x.966.nm, x.960.nm, x.942.nm, x.938.nm, x.926.nm, x.940.nm, x.920.nm, x.978.nm, x.936.nm, x.970.nm
CFS	401	5	x.1150.nm, x.1152.nm, x.1362.nm, x.1364.nm, x.1670.nm
GA	401	302	See Supplementary Information Table 7
SI	401	2	x.1120.nm, x.1264.nm
DBI	401	200	See Supplementary Information Table 7
KMFS	401	18	x.1272.nm, x.1398.nm, x.1662.nm, x.1426.nm, x.1068.nm, x.1436.nm, x.1216.nm, x.1314.nm, x.1560.nm, x.1458.nm, x.974.nm, x.940.nm, x.1126.nm, x.1368.nm, x.1610.nm, x.1490.nm, x.1700.nm, x.1636.nm
BKMFS(50)	401	13	x.1560.nm, x.1368.nm, x.1656.nm, x.1428.nm, x.1046.nm, x.1402.nm, x.1218.nm, x.1320.nm, x.1700.nm, x.1470.nm, x.1090.nm, x.938.nm, x.1628.nm

Table 15. Feature reduction results for all methods on the *Gasoline* dataset.

Method	MAE	RMSE	MAPE
MLR	3.1311	3.7850	0.0358
FRF	3.1366	3.4139	0.0355
RIDGE	0.4883	0.5326	0.0055
LASSO	**0.2083**	**0.2871**	**0.0023**
RFE	1.2629	1.6713	0.0143
CFS	0.7232	0.9333	0.0082
GA	2.0524	2.5547	0.0233
SI	1.1304	1.2514	0.0129
DBI	1.1309	1.2825	0.0128
KMFS	0.3076	0.3870	0.0035
BKMFS(50)	0.2142	0.2909	0.0024

* The bold values indicate the best result.

Table 16. Prediction performance comparison on the *Gasoline* test set based on MAE, RMSE, and MAPE.

4.8 Corn dataset

The Corn dataset is analyzed to evaluate the performance of the proposed method, with the primary objective of predicting the *moisture content* feature. Table 17 presents the number of selected features and the corresponding feature names obtained by each feature selection method, while Table 18 presents the prediction performance of the competing methods on the test dataset. The main result tables report only the best-performing configuration of the proposed method, while the remaining MAE results are given in Table 23. According to the results, the lowest MAE value is achieved by BKMFS(500), with an MAE of 0.1283, indicating the best predictive performance among all competing methods. Among the benchmark methods, LASSO achieves the second-best predictive performance with an MAE of 0.1806, whereas BKMFS(500) reduced the MAE by approximately 28.96%. In terms of RMSE and MAPE, BKMFS(500) also achieves the best predictive performance, further supporting the effectiveness of the proposed method.

Figure 5-(h), presents the sensitivity analysis of the proposed method for the Corn dataset under different bootstrap numbers. The results show that predictive performance improves progressively as the number of bootstrap increased,

ARTICLE IN PRESS

reaches its best performance at 500 iterations. However, a slight deterioration in performance is observed at 1000 bootstrap, suggesting that an excessive number of bootstrap may not provide additional predictive benefits. The incorporation of bootstrap sampling results in an approximately 39.62% improvement in MAE, highlighting its contribution to feature subset quality and predictive performance.

Method	Number of available features	Number of selected features	Selected features
MLR	700	–	–
FRF	700	–	–
RIDGE	700	–	–
LASSO	700	24	V145, V256, V301, V302, V303, V342, V343, V346, V370, V375, V408, V409, V437, V439, V440, V523, V540, V546, V572, V587, V602, V649, V650, V692
RFE	700	9	V53, V46, V41, V27, V24, V35, V52, V54, V51
CFS	700	4	V139, V292, V413, V506
GA	700	574	See Supplementary Information Table 6
SI	700	2	V386, V344
DBI	700	349	See Supplementary Information Table 6
KMFS	700	11	V151, V254, V138, V72, V550, V426, V357, V518, V176, V620, V454
BKMFS(500)	700	30	See Supplementary Information Table 6

Table 17. Feature reduction results for all methods on the *Corn* dataset.

Method	MAE	RMSE	MAPE
MLR	2.1868	2.5850	0.2116
FRF	2.8815	3.7616	0.2787
RIDGE	0.2297	0.2705	0.0226
LASSO	0.1806	0.2351	0.0177
RFE	0.3373	0.3620	0.0333
CFS	0.2252	0.2839	0.0221
GA	2.6363	2.9529	0.2624
SI	0.3336	0.3759	0.0329
DBI	0.5019	0.6339	0.0495
KMFS	0.2125	0.2640	0.0209
BKMFS(500)	**0.1283**	**0.1732**	**0.0129**

* The bold values indicate the best result.

ARTICLE IN PRESS

Table 18. Prediction performance comparison on the *corn* test set based on MAE, RMSE, and MAPE.

4.9 Wine dataset

To assess the effectiveness of the proposed method, the Wine dataset is analyzed, with the primary objective of predicting the *quality* feature. Table 19 presents the number of selected features and the corresponding feature names obtained by each feature selection method, while Table 20 presents the prediction performance of the competing methods on the test dataset. For the proposed method, only the best-performing configuration is reported in the main result tables, and the remaining MAE results are summarized in Table 23. According to the results, the lowest MAE, RMSE, and MAPE values are achieved by KMFS and BKMFS(10), indicating that both methods achieve the same level of predictive performance for the Wine dataset. In addition, MLR, FRF, LASSO, RFE, DBI based k-means, and GA also demonstrate highly competitive predictive performance, with only minor differences across the evaluation metrics.

Figure 5-(i), presents the sensitivity analysis of the proposed method for the Wine dataset under different bootstrap numbers. The results show only minor fluctuations in MAE across bootstrap, indicating that predictive performance is relatively insensitive to the number of bootstrap repetitions for this dataset. This stable behavior further supports the robustness of the proposed feature selection framework.

ARTICLE IN PRESS

Method	Number of available features	Number of selected features	Selected features
MLR	11	–	–
FRF	11	–	–
RIDGE	11	–	–
LASSO	11	10	fixed.acidity, volatile.acidity, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol
RFE	11	10	density, volatile.acidity, pH, sulphates, chlorides, alcohol, residual.sugar, fixed.acidity, citric.acid, free.sulfur.dioxide
CFS	11	1	alcohol
GA	11	8	fixed.acidity, volatile.acidity, residual.sugar, free.sulfur.dioxide, density, pH, sulphates, alcohol
SI	11	7	volatile.acidity, chlorides, residual.sugar, alcohol, pH, total.sulfur.dioxide, fixed.acidity
DBI	11	10	chlorides, density, residual.sugar, volatile.acidity, pH, free.sulfur.dioxide, fixed.acidity, alcohol, citric.acid, total.sulfur.dioxide
KMFS	11	8	chlorides, density, residual.sugar, volatile.acidity, pH, total.sulfur.dioxide, fixed.acidity, alcohol
BKMFS(10)	11	8	chlorides, density, residual.sugar, volatile.acidity, pH, total.sulfur.dioxide, fixed.acidity, alcohol

Table 19. Feature reduction results for all methods on the *Wine* dataset.

Method	MAE	RMSE	MAPE
MLR	0.6175	0.8120	0.1109
FRF	0.6173	0.8113	0.1108
RIDGE	0.6215	0.8161	0.1117
LASSO	0.6189	0.8131	0.1111
RFE	0.6169	0.8119	0.1108
CFS	0.6584	0.8531	0.1182
GA	0.6166	0.8116	0.1107
SI	0.6209	0.8134	0.1113
DBI	0.6193	0.8118	0.1112
KMFS	**0.6155**	**0.8075**	**0.1103**
BKMFS(10)	**0.6155**	**0.8075**	**0.1103**

* The bold values indicate the best result.

Table 20. Prediction performance comparison on the *Wine* test set based on MAE, RMSE, and MAPE.

ARTICLE IN PRESS

4.10 Solubility dataset

Finally, the tenth dataset considered is the Solubility dataset, where the primary objective is to predict the *solubility* of chemical compounds based on their molecular descriptors. Table 21 presents the number of selected features and the corresponding feature names obtained by each feature selection method, while Table 22 presents the prediction performance of the competing methods on the test dataset. For the proposed method, only the best-performing configuration is reported in the main result tables, and the remaining MAE results are summarized in Table 23. According to the results, the lowest MAE value is achieved by BKMFS(250), with an MAE of 0.4743, indicating the best predictive performance among all competing methods. The RMSE results also support this finding, as the proposed method achieves the lowest RMSE value among the competing methods. Ridge and LASSO regression also demonstrate competitive performance, yielding MAE values close to those of the best-performing methods.

According to the number of selected features, RFE retains almost the entire feature set (222 features), whereas CFS selects only 4 features and produces the weakest results. In contrast, the proposed method eliminates approximately half of the original features while achieving the best predictive performance.

Figure 5-(j), presents the sensitivity analysis of the proposed method for the Solubility dataset under different bootstrap numbers. The results show that predictive performance improves progressively up to 250 bootstrap, where the best MAE value is obtained. Beyond this point, a gradual deterioration in performance is observed, suggesting that a moderate to high number of bootstrap repetitions provides the most effective feature selection behavior for this dataset. The incorporation of bootstrap sampling results in an approximately 11.20% improvement in MAE, highlighting its contribution to feature subset quality and predictive performance.

ARTICLE IN PRESS

Method	Number of available features	Number of selected features	Selected features
MLR	228	–	–
FRF	228	–	–
RIDGE	228	–	–
LASSO	228	140	See Supplementary Information Table 9
RFE	228	222	See Supplementary Information Table 9
CFS	228	4	FP015, FP040, FP072, HydrophilicFactor
GA	228	126	See Supplementary Information Table 9
SI	228	100	See Supplementary Information Table 9
DBI	228	111	See Supplementary Information Table 9
KMFS	228	109	See Supplementary Information Table 9
BKMFS(250)	228	113	See Supplementary Information Table 9

Table 21. Feature reduction results for all methods on the *Solubility* dataset.

Method	MAE	RMSE	MAPE
MLR	0.5605	0.7341	0.9867
FRF	0.5605	0.7341	0.9867
RIDGE	0.5067	0.6743	**0.7226**
LASSO	0.5014	0.6585	0.8114
RFE	0.5564	0.7277	0.9061
CFS	1.5765	1.8984	3.8881
GA	0.5470	0.7207	0.8840
SI	0.5385	0.7037	0.9532
DBI	0.4897	0.6436	1.0774
KMFS	0.5341	0.6857	1.1177
BKMFS(250)	**0.4743**	**0.6247**	0.9814

* The bold values indicate the best result.

Table 22. Prediction performance comparison on the *Solubility* test set based on MAE, RMSE, and MAPE.

Methods	Mtcars	Bodyfat	College	Midwest	Automobile	Ethanol	Gasoline	Corn	Wine	Solubility
KMFS	2.3794	4.5601	9.2532	0.1185	2693.8161	1.4570	0.3076	0.2125	0.6155	0.5341
BKMFS(10)	2.1407	4.2379	9.1568	0.1245	2365.4656	1.1953	0.2320	0.2257	0.6155	0.4781
BKMFS(25)	2.3794	4.2025	8.9766	0.1185	2523.3185	1.1064	0.2151	0.1814	0.6203	0.5233
BKMFS(50)	2.1988	4.3788	8.9486	0.1197	2600.6389	1.7871	0.2142	0.1594	0.6203	0.4871
BKMFS(100)	2.3680	4.4129	9.4273	0.1155	2687.0507	1.2080	0.2908	0.1543	0.6155	0.4784
BKMFS(250)	2.1407	4.2525	8.9766	0.1135	2398.1667	0.9681	0.4861	0.1350	0.6203	0.4743
BKMFS(500)	2.3680	4.2002	8.9766	0.1136	2398.1667	1.4073	0.3691	0.1283	0.6155	0.4817
BKMFS(1000)	2.1407	4.2525	8.9766	0.1136	2398.1667	0.9317	0.6163	0.1395	0.6203	0.5088

Table 23. Comparison of MAE based on different bootstrap sampling number for datasets using the proposed method.

ARTICLE IN PRESS

ARTICLE IN PRESS

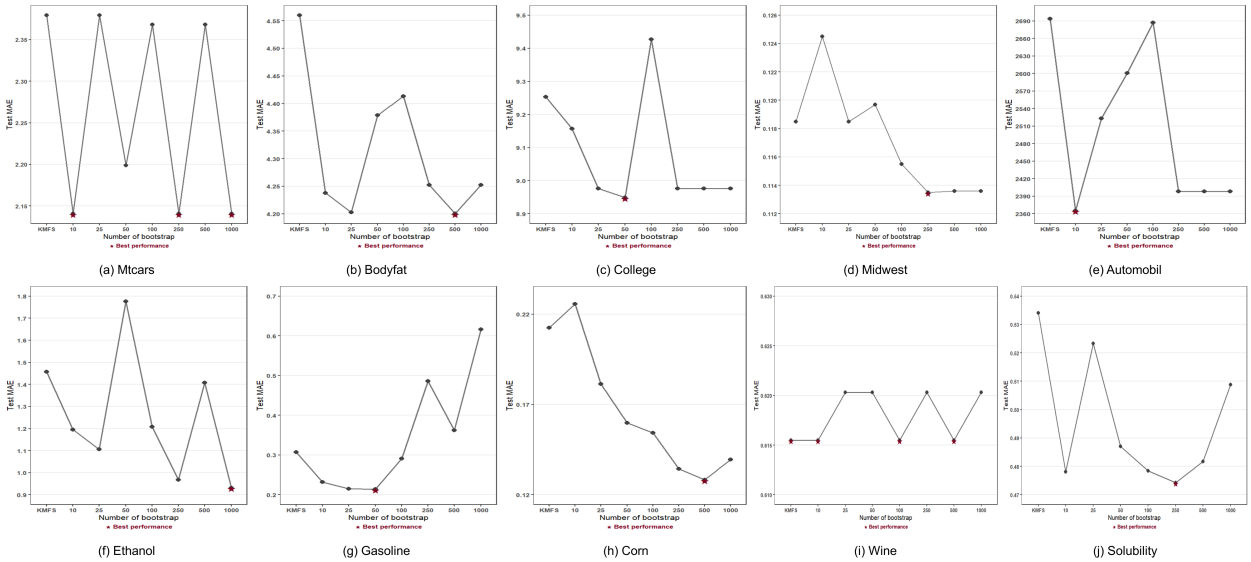

Figure 5. Sensitivity analysis of the proposed method under different bootstrap sizes across all datasets.

4.11 Runtime performance

All analyses are performed using code implemented without parallel computing and are executed independently on two different computer systems: a Dell Inspiron 14 5430 equipped with a 13th Gen Intel® Core™ i5-1335U processor and 16 GB RAM running Microsoft Windows 11, and an ASUS UX310UQ system equipped with an Intel® Core™ i7-7500U processor and 8 GB RAM running Microsoft Windows 10.

To evaluate the computational efficiency of the proposed method, the processing times of the best-performing BKMFS configurations are recorded for each dataset. For each dataset, the bootstrap sampling number corresponding to the lowest MAE value is selected, and the total runtime of the associated feature selection and model training process is measured. Table 24 summarizes the best bootstrap numbers, and the total runtime in seconds. The runtime results reported in Table 24 were obtained on the Dell Inspiron 14 5430 system. The results show that the computational cost varies depending on the instances, and bootstrap numbers.

ARTICLE IN PRESS

Dataset	Best Bootstrap Number	ntest	Runtime (s)
Mtcars	10	3	2.43
Bodyfat	500	25	21.23
College	50	77	36.63
Midwest	250	43	164.60
Automobile	10	20	3.53
Ethanol	1000	16	71.53
Gasoline	50	6	3.21
Corn	500	8	85.86
Wine	10	490	1117.63
Solubility	250	127	65.59

Table 24. Runtime analysis of the best-performing BKMFS models across datasets.

5 Discussion

Empirical findings from ten real-world datasets demonstrate that the proposed feature selection framework provides competitive or superior predictive performance. To better understand the behavior and practical effectiveness of the proposed method, several aspects are discussed in this section, including the relationship between the number of selected features and bootstrap iterations, the overall ranking of the competing methods, the results of the ablation study, the paired comparison test results, and its robustness under different data partitioning ratios.

ARTICLE IN PRESS

Table 3 in Supplementary Information summarizes the optimal bootstrap numbers identified for each dataset together with their corresponding feature numbers. As shown in the table, no clear or monotonic relationship can be observed between the number of available features and the optimal number of bootstrap iterations. For instance, some low-dimensional datasets such as BodyFat require a relatively large number of bootstrap repetitions, whereas some high-dimensional datasets such as Gasoline achieve their best performance with fewer bootstrap iterations. This finding suggests that the optimal bootstrap size is not determined solely by feature dimensionality, but also depends on the intrinsic characteristics of the dataset, such as feature redundancy, and the complexity of the relationship between predictors and the response feature. This relationship could be further explored in future studies using a larger number of datasets with different feature dimensions.

Table 4 in Supplementary Information reports the ranking performance and average rankings of the competing methods in terms of MAE across ten datasets. The results show that the proposed method variants consistently achieve superior ranking performance compared to the benchmark methods. In particular, BKMFS(500) achieves the best overall average ranking (4.35), closely followed by BKMFS(250) (4.50). Among the benchmark methods, LASSO exhibits a competitive performance with achieving the best average ranking (7.2). Overall, the average rankings across the datasets considered in this study suggest that the proposed method framework achieves the highest overall ranking among the competing methods.

Table 25 presents the ablation study results in terms of the percentage MAE improvement achieved by each feature selection method relative to the baseline FRF model. Overall, the bootstrap-based version of the proposed method provides higher performance gains across most datasets compared with the competing feature selection methods, which suggests the contribution of bootstrap sampling to the feature selection process. For example, on the Mtcars dataset, the proposed BKMFS method achieves a 48% improvement in MAE relative to the baseline FRF model. Similar performance patterns can also be observed in the Ethanol, Gasoline, and Corn datasets, where BKMFS yield the highest improvement rates. These findings indicate that bootstrap sampling can contribute to identifying more informative feature subsets and, consequently, improving predictive performance.

Table 5 in the Supplementary Information presents the paired comparison test results between the best-performing BKMFS configuration ($BKMFS_{best}$) and the competing methods across all datasets. At the 95% confidence level, statistically significant differences ($p < 0.05$) are observed in numerous dataset method comparisons, indicating that the proposed method achieves significantly better predictive performance in those cases. More specifically, statistically significant improvements are observed in the College, Midwest, Ethanol, Gasoline, Corn, Wine and Solubility datasets. These findings provide statistical evidence that the proposed method framework improves predictive performance under various data structures and feature numbers.

To further evaluate the robustness and predictive performance of the proposed method against different data partitioning ratios, additional analyses are conducted using alternative train-validation-test split ratios of 70%-15%-15% and 60%-20%-20%, in addition to the 80%-10%-10% setting reported in the experimental results. The detailed results, including MAE, RMSE, MAPE, and the number of selected features, are presented in Supplementary Information Tables 10–19. The purpose of this additional analysis is to investigate whether the predictive performance and feature selection behavior of the proposed method remain stable under different data partitioning conditions. For each alternative partitioning ratio, the complete analysis pipeline, including feature selection, model analyses, and performance evaluation, is repeated from the beginning to ensure a fair and independent assessment. The results show that the proposed framework consistently achieves either the best or highly competitive predictive performance across different split configurations, indicating its robustness to variations in data partitioning. For example, on the mtcars dataset (Supplementary Information Table 10), BKMFS achieves the best MAE values under both alternative split settings (3.1182 and 2.8421). Similarly, for the Midwest dataset (Supplementary Information Table 13), BKMFS again produces the lowest MAE values under both split configurations (0.1095 and 0.1806). In the Ethanol dataset (Supplementary Information Table 15), the proposed method also maintaines the best predictive performance across both partitioning ratios. These findings suggest that the proposed feature selection framework maintains its predictive performance under different train-validation-test partitions.

ARTICLE IN PRESS

ARTICLE IN PRESS

Dataset	SI (%)	DBI (%)	KMFS (%)	BKMFS (%)
Mtcars	29	7	42	**48**
Bodyfat	-12	0	-7	**1**
College	-3	0	-3	0
Midwest	-634	4	1	**5**
Automobile	2	8	-3	**9**
Ethanol	73	92	94	**96**
Gasoline	64	64	90	**93**
Corn	88	83	93	**96**
Wine	-1	0	0	0
Solubility	4	14	5	**15**

Table 25. MAE based ablation study on the effect of feature selection on prediction performance.

6 Conclusion

In conclusion, this study presents a new feature selection approach and evaluates its performance in regression problems using multiple datasets. Compared with widely used feature selection techniques such as CFS, RFE, GA, LASSO, and Ridge regression, as well as alternative cluster number determination criteria based on SI and DBI within the k-means framework, the proposed method demonstrates superior or competitive performance primarily in terms of MAE, with supporting evidence from RMSE and MAPE results.

The contribution of the proposed method lies not only in eliminating less informative features but also overcoming a fundamental limitation of feature selection with k-means clustering. Because k-means initializes cluster centers randomly, classical clustering may lead to unstable or even suboptimal results due to random initialization. In this sense, this study makes two main contributions. First, by integrating bootstrapping with repeated runs, the proposed approach mitigates the effects of randomness, enhances robustness, and improves the probability of identifying feature subsets with superior generalization performance. Second, the method assigns each feature to its nearest cluster center and selects the feature closest to this center as the cluster representative, representing the first application of this strategy in the feature selection literature.

The study highlights the following key points:

- The method introduces novelty by mapping each feature to its nearest cluster center and designating the feature with the smallest distance to the center as the representative.
- By incorporating bootstrap sampling into the k-means clustering algorithm, the method reduces the risk of poor solutions caused by random initialization and improves stability.
- The proposed feature selection method eliminates less informative features carrying similar information in regression models.
- It simplifies the model and enhances interpretability.
- The approach improves predictive performance when integrated with FRF.

Future studies may extend the proposed bootstrap-stabilized feature selection framework in several directions. First, the selected feature subsets may be integrated with alternative regression models, such as LASSO, Ridge regression, support vector regression, neural network based regression methods, to evaluate their predictive performance under different modeling assumptions. Second, the effectiveness of the proposed bootstrap-based stabilization strategy may be investigated using alternative clustering paradigms, such as fuzzy clustering, hierarchical clustering, and spectral clustering, to assess its generalizability beyond hard partitioning methods such as k-means. Finally, applying the proposed framework to large scale and more complex real-world datasets, such as genomic, financial, and industrial data, may provide further insights into its scalability, robustness, and practical applicability.

ARTICLE IN PRESS

ARTICLE IN PRESS

Data availability

The datasets used and/or analysed during the current study (mtcars, college_data, midwest, NIR, gasoline, and corn, solubility) are publicly available R datasets and can be accessed directly within the R environment via their corresponding packages. Specifically, `mtcars` is available in the base R `datasets` package (`data(mtcars)`); `college_data` is available in the `MSMU` package (`library(MSMU); data(college_data)`); `midwest` is available in the `ggplot2` package (`library(ggplot2); data(midwest)`); `NIR` (ethanol) is available in the `chemometrics` package (`library(chemometrics); data("NIR")`); `gasoline` is available in the `pls` package (`library(pls); data("gasoline")`); `corn` is available in the `pcv` package (`library(pcv); data("corn")`) and `solubility` is available in the `AppliedPredictiveModeling` package (`library(AppliedPredictiveModeling); data("solubility")`).

Funding

This study was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) under Grant Number 123F266. The authors thank TÜBİTAK for their support. (Funding: Not received for publication fee)

References

1. Theng, D. & Bhoyar, K. K. Feature selection techniques for machine learning: a survey of more than two decades of research. *Knowl. Inf. Syst.* **66**, 1575–1637, DOI: 10.1007/s10115-023-02010-5 (2024).

2. Buyukkececi, M. & Okur, M. C. A comprehensive review of feature selection and feature selection stability in machine learning. *Gazi Univ. J. Sci.* **36**, 1506–1520, DOI: 10.35378/gujs.993763 (2023).

3. Jović, A., Brkić, K. & Bogunović, N. A review of feature selection methods with applications. In *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*, 1200–1205, DOI: 10.1109/MIPRO.2015.7160458 (Ieee, 2015).

4. Hall, M. A. *Correlation-based feature selection for machine learning*. Ph.D. thesis, The University of Waikato (1999).

5. Chen, X.-W. & Jeong, J. C. Enhanced recursive feature elimination. In *Sixth international conference on machine learning and applications (ICMLA 2007)*, 429–435, DOI: 10.1109/ICMLA.2007.35 (IEEE, 2007).

6. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc. Ser. B: Stat. Methodol.* **58**, 267–288, DOI: 10.1111/j.2517-6161.1996.tb02080.x (1996).

7. Hoerl, A. E. & Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67, DOI: 10.1080/00401706.1970.10488634 (1970).

8. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *J. machine learning research* **3**, 1157–1182 (2003).

9. Guan, D., Yuan, W., Lee, Y.-K., Najeebullah, K. & Rasel, M. K. A review of ensemble learning based feature selection. *IETE Tech. Rev.* **31**, 190–198, DOI: 10.1080/02564602.2014.906859 (2014).

10. Turksen, I. B. Fuzzy functions with lse. *Appl. Soft Comput.* **8**, 1178–1188, DOI: 10.1016/j.asoc.2007.12.004 (2008).

11. Hancer, E., Xue, B. & Zhang, M. A survey on feature selection approaches for clustering. *Artif. intelligence review* **53**, 4519–4545, DOI: 10.1007/s10462-019-09800-w (2020).

12. Alimoussa, M., Porebski, A., Vandenbroucke, N., Thami, R. O. H. & El Fkihi, S. Clustering-based sequential feature selection approach for high dimensional data classification. In *VISIGRAPP (4: VISAPP)*, 122–132, DOI: 10.5220/0010259501220132 (2021).

13. Xu, D. *et al.* Multi-scale supervised clustering-based feature selection for tumor classification and identification of biomarkers and targets on genomic data. *BMC genomics* **21**, 650, DOI: 10.1186/s12864-020-07038-3 (2020).

ARTICLE IN PRESS

ARTICLE IN PRESS

14. Zhu, J., Jang-Jaccard, J., Liu, T. & Zhou, J. Joint spectral clustering based on optimal graph and feature selection. *Neural Process. Lett.* **53**, 257–273, DOI: 10.1007/s11063-020-10383-9 (2021).

15. Wang, P., Xue, B., Liang, J. & Zhang, M. Feature clustering-assisted feature selection approach with differential evolution. *Available at SSRN 4205804* (2023).

16. Mei, M. *et al.* A cooperative hybrid breeding swarm intelligence algorithm for feature selection. *Pattern Recognit.* **169**, 111901 (2026).

17. Tripathi, A. *et al.* Scalable distributed laplacian score for feature selection. In *International Conference on Neural Information Processing*, 1–15 (Springer, 2024).

18. Dwivedi, R., Tiwari, A., Bharill, N., Ratnaparkhe, M. & Tiwari, A. K. A taxonomy of unsupervised feature selection methods including their pros, cons, and challenges: R. dwivedi et al. *The J. Supercomput.* **80**, 24212–24240 (2024).

19. Dwivedi, R. *et al.* A novel chemical property-based, alignment-free scalable feature extraction method for genomic data clustering. *Comput. Electr. Eng.* **123**, 110175 (2025).

20. Dwivedi, R. *et al.* An incremental clustering method based on multiple objectives for dynamic data analysis. *Multimed. Tools Appl.* **83**, 38145–38165 (2024).

21. Ismi, D. P., Panchoo, S. & Murinto, M. K-means clustering based filter feature selection on high dimensional data. *Int. J. Adv. Intell. Informatics* **2**, 38–45, DOI: 10.26555/ijain.v2i1.54 (2016).

22. Sun, Z., Ding, C. & Fan, J. K-means derived unsupervised feature selection using improved admm. *arXiv preprint arXiv:2411.15197* DOI: 10.48550/arXiv.2411.15197 (2024).

23. Boutsidis, C., Zouzias, A., Mahoney, M. W. & Drineas, P. Randomized dimensionality reduction for *k*-means clustering. *IEEE Transactions on Inf. Theory* **61**, 1045–1062, DOI: 10.1109/TIT.2014.2375327 (2014).

24. Tripathi, A., Tiwari, A., Chaudhari, N. S., Ratnaparkhe, M. & Dwivedi, R. Enhancing k-means clustering performance with a two-stage hybrid preprocessing strategy. *Arab. J. for Sci. Eng.* **50**, 15453–15472 (2025).

25. Dwivedi, R., Tiwari, A., Bharill, N. & Ratnaparkhe, M. A novel clustering-based hybrid feature selection approach using ant colony optimization. *Arab. J. for Sci. Eng.* **48**, 10727–10744 (2023).

26. Dwivedi, R., Tiwari, A., Bharill, N. & Ratnaparkhe, M. A hybrid feature selection approach for data clustering based on ant colony optimization. In *International conference on neural information processing*, 659–670 (Springer, 2022).

27. Aladag, C. H., Turksen, I. B., Dalar, A. Z., Egrioglu, E. & Yolcu, U. Application of type-1 fuzzy functions approach for time series forecasting. *Turkish J. Fuzzy Syst.* **5**, 1–9 (2014).

28. Bas, E., Yolcu, U. & Egrioglu, E. Intuitionistic fuzzy time series functions approach for time series forecasting. *Granul. computing* **6**, 619–629, DOI: 10.1007/s41066-020-00220-8 (2021).

29. Tak, N. Type-1 recurrent intuitionistic fuzzy functions for forecasting. *Expert. systems with applications* **140**, 112913, DOI: 10.1016/j.eswa.2019.112913 (2020).

30. Mutlu, E. & Selçuk, G. N. The forecasting of thenumber tourists arriving in turkey with intuitionistic fuzzy regression functions approach. *Turkish J. Forecast.* **8**, 26–32, DOI: 10.34110/forecasting.1512952 (2024).

31. Pehlivan, N. Y. & Turksen, I. B. A novel multiplicative fuzzy regression function with a multiplicative fuzzy clustering algorithm. *Romanian journal information science technology* **24**, 79–98 (2021).

32. Chakravarty, S., Demirhan, H. & Baser, F. Fuzzy regression functions with a noise cluster and the impact of outliers on mainstream machine learning methods in the regression setting. *Appl. Soft Comput.* **96**, 106535, DOI: 10.1016/j.asoc.2020.106535 (2020).

33. Chakravarty, S., Demirhan, H. & Baser, F. Modified fuzzy regression functions with a noise cluster against outlier contamination. *Expert. Syst. with Appl.* **205**, 117717, DOI: 10.1016/j.eswa.2022.117717 (2022).

ARTICLE IN PRESS

ARTICLE IN PRESS

34. Alci, M. & Beyhan, S. Fuzzy functions with function expansion model for nonlinear system identification. *Intell. Autom. & Soft Comput.* **23**, 87–94, DOI: 10.1080/10798587.2015.1136107 (2017).

35. Bodur, M. & Ahmaderaghi, B. A comparison of fuzzy functions with lse and ts-fuzzy methods in modeling uncertain datasets. In *2009 Fifth International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control*, 1–4, DOI: 10.1109/ICSCCW.2009.5379464 (IEEE, 2009).

36. Kizilaslan, B., Egrioglu, E. & Evren, A. A. Intuitionistic fuzzy ridge regression functions. *Commun. Stat. Comput.* **49**, 699–708, DOI: 10.1080/03610918.2019.1626887 (2020).

37. Bas, E., Egrioglu, E., Yolcu, U. & Grosan, C. Type 1 fuzzy function approach based on ridge regression for forecasting. *Granul. Comput.* **4**, 629–637, DOI: 10.1007/s41066-018-0115-4 (2019).

38. Piskin, M. D. & Baş, E. Forecasting monthly housing sales to foreigners with type 1 fuzzy regression functions approach based on ridge regression. *Karadeniz Fen Bilimleri Dergisi* **12**, 571–583, DOI: 10.31466/kfbd.1074832 (2022).

39. Tak, N. & Inan, D. Type-1 fuzzy forecasting functions with elastic net regularization. *Expert. Syst. with Appl.* **199**, 116916, DOI: 10.1016/j.eswa.2022.116916 (2022).

40. Na, S., Xumin, L. & Yong, G. Research on k-means clustering algorithm: An improved k-means clustering algorithm. In *2010 Third International Symposium on intelligent information technology and security informatics*, 63–67, DOI: 10.1109/IITSI.2010.74 (Ieee, 2010).

41. Efron, B. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, 569–593, DOI: https://doi.org/10.1007/978-1-4612-4380-9-41 (Springer, 1992).

42. Thorndike, R. L. Who belongs in the family? *Psychometrika* **18**, 267–276, DOI: 10.1007/BF02289263 (1953).

43. Caliński, T. & Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat. Methods* **3**, 1–27, DOI: doi/abs/10.1080/03610927408827101 (1974).

44. Davies, D. L. & Bouldin, D. W. A cluster separation measure. *IEEE transactions on pattern analysis machine intelligence* 224–227, DOI: 10.1109/TPAMI.1979.4766909 (2009).

45. Friedman, J. *et al.* Package 'glmnet'. *CRAN R Repositary* **595**, 874 (2021).

46. Kuhn, M. Building predictive models in r using the caret package. *J. Stat. Softw.* **28**, 1–26, DOI: 10.18637/jss.v028.i05 (2008).

47. Romanski, P., Kotthoff, L. & Schratz, P. *FSelector: Selecting Attributes* (2023). R package version 0.34.

48. Henderson, H. V. & Velleman, P. F. Building multiple regression models interactively. *Biometrics* **37**, 391–411, DOI: 10.2307/2530428 (1981).

49. Wortley, K. E. *et al.* Absence of ghrelin protects against early-onset obesity. *The J. clinical investigation* **115**, 3573–3578, DOI: 10.1172/JCI26003 (2005). Dataset available at: https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset.

50. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning with Applications in R* (Springer, 2013).

51. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2016).

52. Schlimmer, J. Automobile. UCI Machine Learning Repository, DOI: 10.24432/C5B01C (1985). Dataset available at: https://archive.ics.uci.edu/dataset/10/automobile.

53. Liebmann, B., Friedl, A. & Varmuza, K. Determination of glucose and ethanol in bioethanol production by near infrared spectroscopy and chemometrics. *Anal. Chimica Acta* **642**, 171–178, DOI: https://doi.org/10.1016/j.aca.2008.10.069 (2009).

ARTICLE IN PRESS

ARTICLE IN PRESS

54. Kalivas, J. H. Two data sets of near infrared spectra. *Chemom. Intell. Lab. Syst.* **37**, 255–259, DOI: https://doi.org/10.1016/S0169-7439(97)00038-5 (1997).

55. Eigenvector Research, I. Eigenvector research data sets. https://eigenvector.com/resources/data-sets/. Accessed: 2025-12-01.

56. Cortez, P., Cerdeira, A., Almeida, F., Matos, T. & Reis, J. Wine Quality. UCI Machine Learning Repository (2009). DOI: https://doi.org/10.24432/C56S3T.

57. Tetko, I. V., Tanchuk, V. Y., Kasheva, T. N. & Villa, A. E. Estimation of aqueous solubility of chemical compounds using e-state indices. *J. chemical information computer sciences* **41**, 1488–1493 (2001).

58. Huuskonen, J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J. chemical information computer sciences* **40**, 773–777 (2000).

Author contributions statement

A.U.: methodology, software, validation, formal analysis, investigation, writing—original draft, writing—review and editing. D.Y.: review and editing. N.T.: conceptualization, methodology, software, validation, formal analysis, investigation, writing—original draft, writing—review and editing.

Additional information

Competing interests. On behalf of all authors of the paper, I declare that there is no competing interests.

ARTICLE IN PRESS