Hindawi
International Journal of Intelligent Systems
Volume 2023, Article ID 2400194, 16 pages
https://doi.org/10.1155/2023/2400194

Research Article

A Model-Free Feature Selection Technique of Feature Screening and Random Forest-Based Recursive Feature Elimination

Siwei Xia [1] **and Yuehan Yang** [2]

[1] *School of Science, Civil Aviation Flight University of China, Deyang, China*
[2] *School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, China*

Correspondence should be addressed to Yuehan Yang; yyh@cufe.edu.cn

Received 19 April 2023; Revised 7 August 2023; Accepted 14 August 2023; Published 29 August 2023

Academic Editor: Mohammad R. Khosravi

This paper studies data with mass features, commonly observed in applications such as text classification and medical diagnosis. We allow data to have several structures without requiring a specific model and propose an efficient model-free feature selection procedure. The proposed method can work with various types of datasets. We demonstrate that this method has several desirable properties, including high accuracy, model-free, and computational efficiency and can be applied to practical problems with different modelings. We prove that the proposed method achieves selection consistency and L_2 consistency under mild regularity conditions. We conduct simulations on various datasets, including data generated from the generalized linear model, additive model, Poisson regression, and binary classification model. These simulations illustrate the superior performance of the proposed method compared to other existing methods across different model settings. In addition, we apply our method to two real examples, the Tecator dataset and the Daily Demand Orders dataset, both of which are continuous and high dimensional. In both cases, our method consistently achieves high accuracy in prediction and model selection.

1. Introduction

Due to the rapid development of data technology, feature selection is a critical component in both statistics and machine learning. High-dimensional and ultrahigh dimensional datasets are commonly encountered in various fields, including finance, text classification, biology, and medicine [1–6]. In the case of financial stock market analysis, the samples correspond to the latest trading days, and the features represent the returns of a large number of stocks. The number of samples is limited, while the number of features is often much higher than the number of samples [7]. The presence of numerous redundant features can weaken model generalization and make data analysis more challenging [8]. The efficiency of feature selection is crucial, as it focuses on identifying a small subset of informative features that contain the necessary information to address specific concerns arising from a study. In many data analyses, feature selection is a significant and frequently used dimensionality reduction technique and is often regarded as a key preprocessing step in data analysis that offers advantages such as interpretability, accuracy, lower computational costs, and reduced risk of overfitting [9, 10].

1.1. Literature Review. There has been a considerable amount of research on feature selection, which can generally be categorized into three types: embedded, filter, and wrapper methods [11, 12]. Embedded approaches involve model learning by selecting variables during the learning process using methods such as objective function optimization, change calculation, and selecting the set of variables with the best solution as the best model. Lasso [13] with the l_1 regularization penalty and decision trees [14] are two typical examples of embedded methods. Penalized regularizations that shrink estimates by penalty functions, such as

Lasso, SCAD [15], MCP [16], and LSP [17] have been extensively studied for high-dimensional data. These methods estimate and select features simultaneously and are computationally efficient. Many researchers have studied their algorithms and statistical properties, such as the least angle regression [18], coordinate descent algorithm [19, 20], iterative majority minimization [21], and theoretical guarantees for Lasso [22]. However, regularization methods are restricted by model assumptions and are mostly applied to regression problems. Furthermore, the selection of tuning parameters can impact the estimation accuracy and computational cost.

Filter techniques, also known as variable ranking techniques, involve calculating a specific statistical measure for each variable and ranking the features based on this measure. They select the optimal subset of features according to predetermined selection criteria. These techniques are often used as preselection strategies that are independent of the machine learning algorithms applied later in the analysis [23]. Since they do not rely on inductive algorithms, they are practically free. Classic ranking criteria such as Fisher score [24] and Pearson correlation [25] are commonly used in filter techniques. In addition, nonlinear approaches such as Joint mutual information maximization and normalized joint mutual information maximization [26] use mutual information and the maximum of the minimum criterion to balance accuracy and stability. Ramírez-Gallego et al. [27] proposed fast-mRMR, an extension of the mRMR filter method based on several optimizations and can tackle high-dimensional big data. Moreover, F-score technique is a valuable filter for binary datasets, which has been found successful applications in numerous biomedical contexts [28–31]. The key characteristics of filter techniques are their speed, simplicity, and efficiency [32, 33].

Feature screening is a type of filter technique that addresses the problem of ultrahigh dimensionality by screening out irrelevant variables. Unlike other feature selection methods that aim to identify a subset of informative features, feature screening is less ambitious as it only aims to discover a majority of irrelevant variables. In other words, it identifies a set of features that contains important variables while allowing many irrelevant variables to be included.

The concept of feature screening as a filter technique is essential in solving problems caused by ultrahigh dimensionality. Fan and Lv [34] proposed this idea for the first time through a feature screening method called sure independence screening (SIS). The paper aimed to remove redundant features by ranking their marginal Pearson correlations and provided theoretical results called the sure independence screening property. These results showed that the remaining feature set contains all the important variables with high probability. SIS has gained popularity among ultrahigh dimensional analyses due to its facility, effectiveness, and promising numerical performance [35, 36]. Feature screening has since been applied to many problems, including parametric models (e.g., [37–39]) and semi-parametric or nonparametric models (e.g., [40–44]). The main drawback of this filter technique is that the selection process does not take into consideration the performance of the learning model. The previous studies mentioned above also have model limitations and cannot accurately select the active set.

The last category is wrapper techniques, which involves searching for the optimal model by computing the model performance for every possible combination of available features, similar to a search problem. The goal is to select the best model with the highest performance. Wrappers are widely studied for their simplicity, availability, and generalizability. Commonly used wrapper methods include forward selection-based approaches [42, 45] and backward selection-based approaches [46, 47]. However, these methods can be computationally expensive and are not suitable for ultrahigh dimensional data. To address these challenges, researchers have proposed advanced methods such as the forward-backward selection with early dropping [48], sequential conditioning approach [49], and forward variable selection procedures for ultrahigh dimensional generalized varying coefficient models [50]. Although these methods are suitable for ultrahigh dimensionality, they still rely on model-based feature selection procedures.

1.2. Motivation and Contribution. Based on the existing results, we summarize that an appealing feature selection approach should satisfy the following three properties:

(i) High accuracy, which means that the subset consisting of informative features can be correctly selected. This is a basic requirement, and most methods have desirable accuracy under suitable conditions.

(ii) Model-free, i.e., it can be implemented without requiring a specific model. Specifying a model is challenging for empirical analysis. Recently, the model-free feature selection method has become a hot research topic for its generalization and validity.

(iii) Computational efficiency, especially for the ultrahigh dimensional dataset that is usually time-consuming.

For the second property, model-free feature screening is first proposed by Zhu et al. [51]. After that, He et al. [52] proposed a quantile-adaptive model-free feature screening framework for high dimensional heterogeneous data. Mai and Zou [53] further developed the fused Kolmogorov filter for model-free feature screening with categorical, discrete, and continuous responses. Liu et al. [54] proposed a model-free and data-adaptive feature screening method named PC-Screen, which is based on ranking the projection correlations between features and response. A state-of-the-art approach to wrapper methods without model restrictions is recursive feature elimination (RFE), a sequential backward elimination, i.e., support vector machine-based recursive feature elimination [55–57], random forest-based recursive feature elimination [58, 59], partial least squares-based recursive feature elimination [60]. Motivated by RFE, Xia and Yang [61] proposed an iterative model-free feature screening procedure named forward recursive selection.

Regarding the third property, filter-based feature selection methods often have lower computational complexity than embedded and wrapper techniques [59, 62]. Some improvements have been made to the computational efficiency of wrapper methods. For example, Borboudakis and Tsamardinos [48] introduced early dropping to increase computational efficiency. Honda and Lin [50] and Xia and Yang [61] reduced computational consumption by adding a stopping rule that takes into account the model size.

Since the abovementioned approaches always cannot satisfy the three properties simultaneously, to fill this gap, this paper proposes a model-free feature selection procedure for ultrahigh dimensional datasets. The proposed approach, FK-RFE, combines the fused Kolmogorov filter and random forest-based recursive feature elimination techniques to overcome model limitations and reduce computational complexity. The approach consists of two phases: the first phase ranks the features based on their relevance and retains the most relevant features based on a threshold value; the second phase evaluates successive subsets of features according to a predefined search strategy and an optimality criterion. Both theoretically and empirically, we demonstrate the effectiveness of the proposed method in addressing the challenges associated with ultrahigh dimensional datasets. We show that the proposed method exhibits desirable properties: model-free, high accuracy, and computational efficiency. The specific contributions are shown in the following points:

(1) The first contribution is that the proposed model-free approach can be applied to a variety of ultrahigh dimensional datasets. Specifically, we propose to use the fused Kolmogorov filter and random forest to remove model assumptions and data assumptions. Besides, this approach combines the advantages of the wrapper and filter strategies and is computationally efficient, making it well-suited for datasets with large numbers of features. We demonstrate that our method is capable of handling ultrahigh dimensional data with complex structures.

(2) We address the challenge of the theoretical guarantees for model-free algorithms and prove the convergence of the proposed algorithm. Specifically, we prove that the feature selection procedure is selection consistent and L_2 consistent under mild conditions. This theoretical analysis provides a solid foundation for the effectiveness of the proposed method and further validates its suitability for ultrahigh dimensional datasets.

(3) We evaluate the performance of our proposed method against several existing methods in various models, including the generalized linear model, additive model, and Poisson regression model, et al., in high and ultrahigh dimensional settings. We conduct simulations and apply the proposed approach to two real datasets. Our experimental results demonstrate the effectiveness and efficiency of our proposed method.

The remainder of this paper is organized as follows. Section 2 describes the proposed method, the algorithm, and its advantages. Section 3 illustrates the theoretical properties. Sections 4 and 5 present the simulation and application results. Section 6 concludes the paper. Technical details are provided in Appendix.

2. Methods

In this section, we introduce the proposed model-free feature selection procedure, FK-RFE. This method incorporates a filter phase and wrapper phase possessing the advantages of feature screening, recursive feature elimination, and random forest. In the following, we show that this technique is efficient and can be applied to various data. For simplicity of description, we first consider a supervised problem with a response Y, predictors $X = (X_1, \ldots, X_p)$ and the following model framework:

$$Y = f(X) + \epsilon, \tag{1}$$

where f is a measurable function and can be any model, e.g., parametric, semiparametric, or nonparametric model. ϵ, a noise term, is independent of predictor X_j with $E(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2 \in (0, \infty)$. When the dimension p becomes very large, a reasonable requirement is the sparsity assumption that only a small subset of variables is responsible for modeling Y. Before presenting the complete algorithm of the proposed procedure, we first introduce the two phases that comprise the algorithm. Furthermore, we will delve into the two fundamental techniques used in each respective phase: the fused Kolmogorov filter and the random forest.

2.1. The First Phase of FK-RFE: Filter.

2.1. The First Phase of FK-RFE: Filter. In this section, we introduce the first screening phase of our proposed method, the fused Kolmogorov filter, which was originally introduced for model-free feature screening [53]. The fused Kolmogorov filter enjoys the sure screening property under weak regularity conditions and is a powerful technique for datasets with strongly dependent covariates. We extend this technique to handle a variety of datasets, including both parametric and nonparametric regression. In our proposed algorithm, we compute the fused Kolmogorov filter statistics for all the features and then select the top d_n features based on these values. This first phase serves as a rapid downscaling step, where the parameter d_n is a predetermined positive integer. In this part, we introduce the definition and the calculation of the fused Kolmogorov filter statistic.

Consider a dataset with n samples pairs denoted as (x_{ij}, y_i) obtained from the response Y and predictors $X = (X_1, \ldots, X_p)$ $(i = 1, \ldots, n,\ j = 1, \ldots, p)$. The main idea of this filter is that X_j and Y are independent if and only if the conditional distributions of X_j given different values of Y remain the same. Thus, the fused Kolmogorov filter focuses on a difference measure denoted as K_j for X_j and Y,

$$K_j = \sup_{y_1, y_2} \sup_x \left| F_j(x \mid Y = y_1) - F_j(x \mid Y = y_2) \right|, \tag{2}$$

where F_j denotes the generic cumulative distribution function (CDF) of X_j. Based on the definition, $K_j = 0$ if and only if X_j is independent of Y. Estimating K_j is straightforward for the binary response case; for instance, when $Y = 1, 2$, we have

$$\widehat{K}_j = \sup_x \left|\widehat{F}_j(x \mid Y = 1) - \widehat{F}_j(x \mid Y = 2)\right|, \tag{3}$$

where $\widehat{F}_j$ denotes the generic empirical CDF. If Y is continuous, following the approach suggested by [53], the approximation of K_j involves partitioning the response. Specifically, define $N (\in \mathbb{N}^+)$ distinct partitions of the response values. Let G_t denote the t th partition, consisting of g_t slices $(t = 1, \ldots, N)$, i.e.,

$$G_t = \left\{[a_{l-1}, a_l): a_{l-1} < a_l \text{ for } l = 1, \ldots, g_t, \text{and} \cup_{l=1}^{g_t} [a_{l-1}, a_l) = \mathbb{R}\right\}, \tag{4}$$

where each $[a_{l-1}, a_l)$ denotes a slice, $a_0 = -\infty, a_{g_t} = +\infty$, and the interval $[a_0, a_1) = (-\infty, a_1)$. Then, we define a random variable $H_j \in \{1, \ldots, g_t\}$ such that $H_j = l$ if Y falls into the lth slice. The partition G_t should consist of intervals bounded by the $(1/g_t)$ th sample quantiles of Y. For a given partition G_t, $K_j^{G_t}$, which is the approximation of K_j, is defined as follows:

$$K_j^{G_t} = \max_{l,\gamma} \ \sup_x \left|F_j\left(x \mid H_j = l\right) - F_j\left(x \mid H_j = \gamma\right)\right|. \tag{5}$$

Note that F_j represents the generic CDF of X_j. We have $F_j(x \mid H_j = l) = P(X_j \leqslant x \mid H_j = l)$, where $l = 1, \ldots, g_t$. Naturally, the empirical version of $K_j^{G_t}$ based on the samples (x_{ij}, y_i) is defined as follows:

$$\widehat{K}_j^{G_t} = \max_{l,\gamma} \ \max_{l,\gamma} \left|\widehat{F}_j\left(x \mid H_j = l\right) - \widehat{F}_j\left(x \mid H_j = \gamma\right)\right|, \tag{6}$$

where $\widehat{F}_j$ denotes the generic empirical CDF, defined as $\widehat{F}_j(x \mid H_j = l) = (1/n_l)\sum_{\{H_j = l\}} 1(x_{ij} \leqslant x)$, where n_l is the sample size of $\left\{H_j = l\right\}$. The fused Kolmogorov filter statistic is then computed as the sum over N different partitions, which integrates various slicing schemes, and is defined as follows:

$$\widehat{K}_j = \sum_{t=1}^{N} \widehat{K}_j^{G_t}. \tag{7}$$

In practice, utilizing different partitioning strategies for the response does not significantly affect feature screening results. We choose that $g_t \leqslant \lceil \log n \rceil$ for all t so that each slice contains a sufficient sample size for all slicing strategies. In addition, if Y is a multilevel categorical variable, such as $Y = 1, \ldots, g$, a single partition is used $(N = 1)$. This partition, denoted as G, is directly derived from Y's level, i.e., $G = \{1, \ldots, g\}$. In this case, we simply set $H = Y$.

For ease of notation, we denote the fused Kolmogorov filter as $\widehat{K}_j$ to represent its form across all data types. Consequently, the screening set of the fused Kolmogorov filter, denoted as V_0, is defined as follows:

$$V_0 = \left\{1 \leqslant j \leqslant p: \widehat{K}_j \text{ is among the first } d_n \text{ largest of all}\right\}. \tag{8}$$

2.2. The Second Phase of FK-RFE: Wrapper. In the second half of our proposed method, we utilize a backward strategy known as recursive feature elimination. At each step of this strategy, the variable importance ranking is updated under the current model, and the feature with the lowest importance measure is removed from the active set. This strategy was introduced by Guyon et al. [55] for support vector machines and has gained popularity in numerous fields, such as gene selection [63, 64] and medical diagnosis [65, 66].

Random forest is an ensemble learning approach that operates on the bagging method's mechanism [67]. It consists of a collection of decision trees. Let $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be the sample set of (X, Y). $\widehat{f}$ is an estimate of f used to predict Y. The trees are constructed using M bootstrap samples $D_1, \ldots, D_M$ of D. The learning rule of random forest is the aggregation of all the tree-based estimators denoted by $\widehat{f}_1, \ldots, \widehat{f}_M$ where the aggregation is calculated based on the average of the predictions $\widehat{f} = 1/M\sum_{m=1}^{M}\widehat{f}_m$.

Random forest accesses the relevance of a predictor by the permutation importance measure [58, 67], which is used to eliminate features in the wrapper phase. This measure is based on the idea that a variable X_j is relevant to Y if the prediction error increases when we break the link between X_j and Y, and this link can be broken by random permuting the observations of X_j. That is, for $j = 1, \ldots, p$, set $X_{(j)} = (X_1, \ldots, X_j', \ldots, X_p)$ be the random vector in which X_j' is an independent replication of X_j. The permutation importance measure is given by

$$I\left(X_j\right) = E\left[\left(Y - f\left(X_{(j)}\right)\right)^2\right] - E\left[(Y - f(X))^2\right]. \tag{9}$$

Note that the random permutation also breaks the link between X_j and other predictors. In other words, X_j' is independent of Y and other predictors $X_{j'}$, $j' \neq j$, simultaneously. Denote $\overline{D}_m = D/D_m$ as the out-of-bag samples of D_m to contain the observations which are not selected in D_m. Let $\overline{D}_m^j$ be the permuted out-of-bag samples by random permutations of the observations of X_j. The empirical permutation importance measure is expressed as follows:

$$\widehat{I}\left(X_j\right) = \frac{1}{M}\sum_{m=1}^{M}\left[R\left(\widehat{f}_m, \overline{D}_m^j\right) - R\left(\widehat{f}_m, \overline{D}_m\right)\right], \tag{10}$$

where $R(\widehat{f}_m, T) = (1/|T|)\sum_{i:(x_i, y_i)\in T}(y_i - \widehat{f}_m(x_i))^2$ for sample set $T = \overline{D}_m^j$ or $T = \overline{D}_m$. The permutation importance measure is recalculated to rank the predictors in each iteration. In addition to other criteria, the permutation importance measure has proven to be effective for leading variable selection methods [58, 68].

The random forest has several advantages. First, it allows us to deal with different data types, including both continuous and categorical variables. Second, both theoretical and empirical evidence support the application of this method. In addition, combining with the permutation importance, we achieve high accuracy in feature selection and outperform other compared methods.

2.3. FK-RFE. In this section, we present the FK-RFE in detail. The proposed algorithm consists of a filter phase and a wrapper phase. In the filter phase, we use the fused Kolmogorov filter, a feature screening technique, to remove a large number of uninformative features and obtain a reduced active set V_0, which includes the true model. Subsequently, in the wrapper phase, we utilize the random forest to train the model and rank the features based on their permutation importance measure. During this step, we iteratively update the active set by eliminating the least significant feature. In each iteration, we rerank the remaining features by recalculating the permutation importance measure, as it is more effective than the approach without reranking [69, 70]. We determine the optimal subset of features based on the best model performance. The pseudocode of FK-RFE with the execution process is given in the following Algorithm 1, and the flowchart is given in Figure 1.

We utilize the fused Kolmogorov filter as the first screening phase in our proposed method, which has several main advantages. First, it allows the method to be widely applicable to various types of data by being free from model restrictions. Second, it is fast and straightforward, especially for ultrahigh dimensional settings. Third, it has theoretical guarantees, as we show in the next section, that the subset obtained from FK-RFE includes all relevant variables. Finally, it achieves screening efficiency, meaning that after the screening phase, the model size is controlled by d_n.

The random forest-based recursive feature elimination and permutation importance measure have several advantages. The first advantage is that they allow the proposed method to apply to different types of data, including both continuous and categorical variables. The second advantage is that the proposed method achieves high accuracy in feature selection, as supported by both theoretical guarantees and empirical evidence. In particular, the permutation importance measure provides a reliable and robust way to rank the importance of features, and the recursive feature elimination algorithm can iteratively eliminate unimportant features, leading to a final subset of relevant features.

The proposed algorithm contains some parameters, hyperparameters, and criteria. To obtain the optimal set, we utilize the mean squared error (MSE) for continuous response or the out-of-bag (OOB) error for the multilevel categorical response as the criteria for model performance. The first screening phase, known as the filter, involves certain related parameters. First, we use a parameter denoted as d_n to control the number of features selected through the fused Kolmogorov filter statistics. For ultrahigh dimensional sparse models, we follow the common setting that $d_n = a\lceil n/\log n\rceil$, where a is a given constant [34, 35, 53]. In this case, the first screening phase tends to choose a larger model size compared to the true size of the relevant features. In the wrapper phase, there are some hyperparameters in random forest, such as the number of trees to grow and the number of variables randomly sampled as candidates at each split. According to our numerical experience and the recommendation from references [58, 60, 71], the results do not different much over a range of hyperparameters. Thus, we follow the regular setting of the random forest, as recommended by references, that apply the default hyperparameters provided by the R package random forest. For example, the number of trees to grow is set at 500. For the number of variables randomly sampled as candidates at each split, the values are $\sqrt{p}$ for classification and $p/3$ for regression. The minimum size of terminal nodes is set at 1 and 5 for classification and regression, respectively. More details can be found in [72]. The source codes of the proposed algorithm and datasets are available on GitHub (URL: https://github.com/momoxia1992/FK-RFE).

2.4. Discussions and Comparison with Other Methods and Algorithms. In this section, we aim to discuss the characteristics of the proposed method and compare it with other existing methods. Notably, the FK-RFE approach does not require any model or data assumptions, rendering it suitable for diverse datasets and applicable to nonparametric, semiparametric, and parametric scenarios. Moreover, by employing random forest for training, the proposed method is robust to noise, missing data, outliers, and ultrahigh dimensional data. We substantiate these claims through numerical experiments. To gain further insight into the FK-RFE, we differentiate it from other methods based on the following criteria. We first compare it with some specific methods:

(i) Compared to forward selection [73], which starts with an empty set and adds features one by one, FK-RFE uses a backward strategy, which starts with a large set of features and removes the least important ones iteratively. This strategy efficiently reduces the risk of overfitting, providing better adaptation to noisy or redundant features and improving the generalization ability of the model.

(ii) Compared to Lasso and other regularization methods that are commonly solved using the coordinate descent algorithm, FK-RFE offers many advantages. It does not rely on tuning penalties or model assumptions, for example, linearity or normality assumptions. In this case, FK-RFE would have better performance in situations where the model assumptions fail and avoids the effects of the selection of tuning parameters.

(iii) Compared to filter methods, such as mutual information-based methods [26, 74], FK-RFE considers the joint effects of features by using the fused Kolmogorov filter, which is more suitable for situations where covariates are strongly dependent on each other. Furthermore, the proposed method is applicable to a wide range of data types. This broad applicability makes it more suitable for practical problem compared to some methods designed specifically for certain data types, such as F-score [28] for binary data and SIS [34] for continuous data.

(iv) Compared to wrapper methods, such as sequential forward selection [75], FK-RFE is computationally efficient, especially for ultrahigh dimensional data,

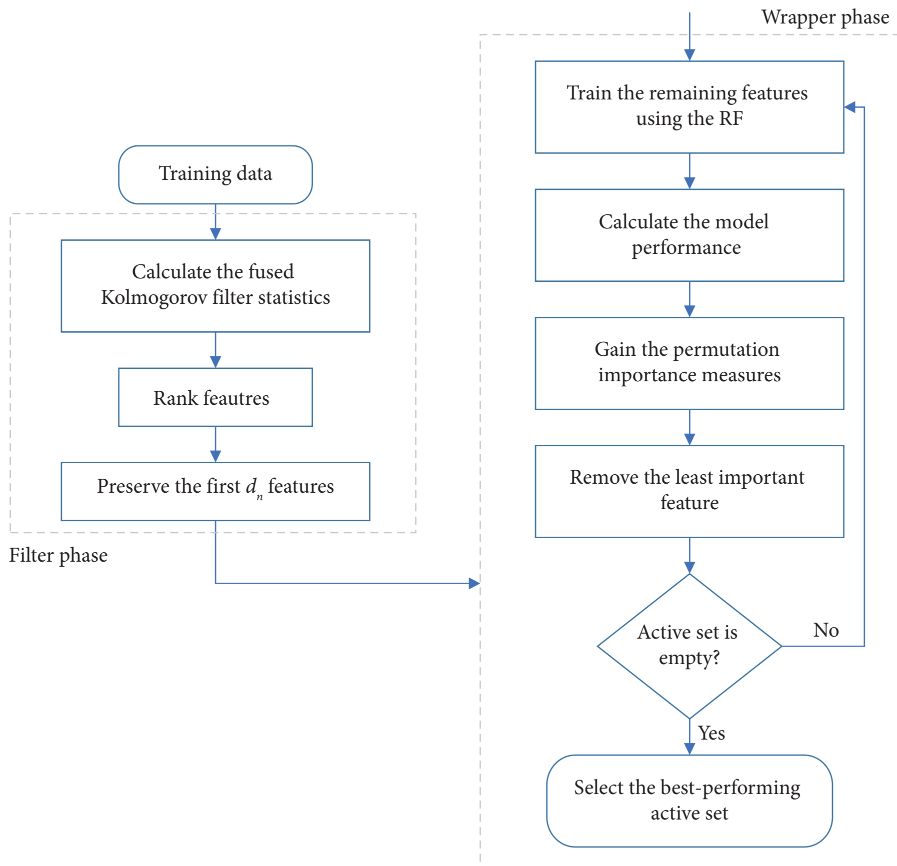

FIGURE 1: The flowchart of FK-RFE consists of two main phases: the filter phase and the wrapper phase. The filter phase aims to efficiently downscale the features, reducing computational consumption. On the other hand, the wrapper phase focuses on selecting the most appropriate subset to ensure accuracy.

Inputs:
X: Sample of predictors
Y: Sample of response
M: Number of trees
d_n: Threshold of filter
Output: The set with the best model performance
Filter phase:
(1) Rank the features according to the fused Kolmogorov filter statistic $\widehat{K}_j$, $j = 1, \ldots, p$
(2) Obtain the reduced set
$V_0 = \left\{j: \widehat{K}_j \text{ is among the first } d_n \text{ largest of all}\right\}$
Wrapper phase:
(1) **for** all the remaining features **do**
(2) Train the model using the random forest
(3) Calculate the model performance
(4) Calculate the permutation importance measures
(5) Update V_{l-1} to V_l by eliminating the least important feature
(6) Update $l = l + 1$
(7) Continue until no features are left
(8) **end for**

ALGORITHM 1: FK-RFE.

because it reduces the number of features in the wrapper phase by screening out irrelevant features in the filter phase.

(v) Compared to other similar iterative algorithms, the FK-RFE method also offers some advantages. For instance, recursive feature elimination [58] is computationally demanding, as the iteration starts with all variables and ends when no variables remain. Forward recursive selection [61] is suitable for high-dimensional data, but the number of iterations is determined by the number of samples, leading to high computational costs when dealing with large datasets.

The FK-RFE method offers a unique combination of the wrapper and filter techniques without their respective disadvantages. In the first phase, it efficiently reduces dimensions using the filter technique, while in the second phase, it avoids the computational burden of the wrapper method. This combination leads to high accuracy in feature selection. Unlike model-based techniques such as regularization approaches [13, 15, 76] and model-based forward selections [49, 73, 77], the proposed FK-RFE method is model-free and requires fewer assumptions, making it suitable for a wider variety of data formats. Moreover, the algorithm requires only one parameter d_n, which is not crucial and can be easily calculated without cross-validation, BIC, or other parameter selection techniques.

3. Consistency Analysis

Noted that we consider sparse learning problems with the model framework (1), i.e.,

$$Y = f(X) + \epsilon. \tag{11}$$

Considers a subset $S = \{j\colon X_j \text{ is relevant to } Y\} \subset \{1, 2, \ldots, p\}$ with cardinality $|S| = q$ that much less than the dimension p. Formally speaking, we refer to a predictor $j \in S$ as informative or relevant. If a predictor j belongs to the complement of S, i.e., $j \in S^c = \{1, \ldots, p\}\backslash S$, it is regarded as uninformative or unimportant. In practice, a forest can only be created with a finite number of trees. On the other hand, in the theoretical analysis, it is generally assumed that M tends to infinity. This is because when $M = \infty$, the predictors do not depend on the realization of the specific tree in the forest. To simplify the proof, we follow this assumption and consider the consistent property among the infinite forest. The infinite forest estimate is defined by $\overline{f} = E(\widehat{f})$. By the law of large numbers,

$$\overline{f} = \lim_{M \longrightarrow \infty} \widehat{f} = \lim_{M \longrightarrow \infty} \frac{1}{M} \sum_{m=1}^{M} \widehat{f}_m, \tag{12}$$

where more details can be found in [67, 78]. We consider the following regularity conditions, under which we can guarantee the convergence of Algorithm 1, that the FK-RFE is selection consistent and L_2 consistent.

C1. There exists a set S_1 such that $S \subset S_1$ and

$$\triangle_{S_1} = \min_t \left(\min_{j \in S_1} K_j^{G_t^o} - \max_{j \notin S_1} K_j^{G_t^o} \right) > 0, \tag{13}$$

where $K_j^{G_t^o}$ is $K_j^{G_t}$ under oracle partition G_t which contains the intervals bounded by the $1/g_t$ th theoretical quantiles of Y.

C2. For any b_1, b_2 such that $P(Y \in [b_1, b_2)) \leqslant 2/\min\{g_t\}$, we have

$$\left| F_j(x \mid y_1) - F_j(x \mid y_2) \right| \leqslant \frac{\triangle_{S_1}}{8}, \tag{14}$$

for all x, j and $y_1, y_2 \in [b_1, b_2)$.

Theorem 1. *Suppose conditions C1-C2 hold. Assume the importance measure $\widehat{I}(X_j)$ is an unbiased estimator of $I(X_j)$, i.e., $\lim_{M \longrightarrow \infty} E(\widehat{I}(X_j)) = I(X_j)$, as $n \longrightarrow \infty$, and the infinite random forest is L_2 consistent. Then the FK-RFE is selection consistent. That is, denoting $\widehat{S}$ to be the set selected by the FK-RFE, we have*

$$P(\widehat{S} = S) \longrightarrow 1, \text{as}\, n \longrightarrow \infty. \tag{15}$$

Remark 2. Conditions C1 and C2 follow the conditions C1 and C2 in [53], which guarantee the sure screening property of the fused Kolmogorov filter. Specifically, Condition C1 ensures that the predictors in the set S are marginally important, which is a regular condition in marginal screening approaches. Condition C2 guarantees that the sample quantiles of Y are close enough to the population quantiles of Y. Both conditions are mild.

Remark 3. The validity of the importance measure is formally proven to be valid under some general assumptions in [68]. This guarantees that the permutation importance measure of the informative predictor converges to a nonzero constant and that one of the uninformative predictors converges to 0 with probability. Therefore, the uninformative predictors are eliminated before the informative ones. The permutation importance measure is widely studied in various references, such as [58, 61, 68, 79]. For instance, Gregorutti et al. [58] proposed $I(X_j) = 2\mathrm{Var}(f_j(X_j))$ under an additive regression model, i.e., $f(X) = \sum_{j=1}^{p} f_j(X_j)$. Furthermore, Ramosaj and Pauly [68] proved that under more general assumptions and model (1), $I(X_j)$ equals $E[(f(X) - f(X_{(j)}))^2]$ for $j \in S$, or equals 0 for $j \in S^c$.

Remark 4. It is worth noting that another important requirement for Theorem 1 is the L_2 consistency of the random forest estimator. This requirement has been extensively studied in the literature, with numerous references providing insights into this topic. For instance, Breiman [67] established an upper bound on the generalization error of forests based on the correlation and strength of individual trees. Denil et al. [80] proved the consistency of online random forests, while Scornet et al. [78] demonstrated L_2 consistency of random forests in an additive regression framework. Athey et al. [81] proposed a generalized random forest and developed an asymptotic and consistency theory for it. Given the vastness of this topic, we refer interested readers to the aforementioned references for a more detailed discussion of L_2 consistency in random forests and state the following result without proof.

Proposition 5. *Assume the infinite random forest is L_2 consistent. The FK-RFE is L_2 consistent too.*

4. Simulations

In this section, we conduct a comparative analysis of the FK-RFE with other feature selection methods on simulated datasets spanning from low to ultrahigh dimensional settings. The sample size is fixed at $n = 100$, while the number of features varies from $p = 100$ to $p = 2000$. As recommended by [53], we set $g_t = 3, 4$ in the fused Kolmogorov filter of the FK-RFE for $\lceil \log n \rceil = 4$, and the threshold $d_n = \lceil n/\log n \rceil$. We compare the FK-RFE with five other feature selection methods, namely, recursive feature elimination (RFE) [58], forward recursive selection (FRS) [61], Lasso [13], forward-backward selection with early dropping (FBED) [48], and F-score [28]. Note that the F-score is proposed for binary data, thus we only apply and compare this method in this case. Lasso is the corresponding form under logistic regression when the response is binary. We consider six models in this simulation study.

Example 1. $Y = \exp(X_1 + X_2 + X_3 + X_4 + X_5) + \epsilon$.

Example 2. $Y^{1/9} = 2.8X_1 - 2.8X_2 + \epsilon$.

Example 3. $Y = (X_1 + X_2 + 1)^3 + \epsilon$.

Example 4. $Y = 2(X_1 + X_2) + 2\tan(\pi X_3/2) + 5X_4 + \epsilon$.

Example 5. $Y \sim \text{Poisson}(u)$, where $u = \exp(0.8X_1 - 0.8X_2)$, $X_j \sim t_2$ independently.

Example 6. $Y \sim B(1, \pi)$, where $\ln(\pi/1 - \pi) = 3X_1X_2 + 2X_3 + 2X_4$.

The error $\epsilon \sim N(0, 1)$. In the Examples 1–4 and 6, the predictor $X = (X_1, \ldots, X_p) \sim N(0, \Sigma)$ and the covariance matrix is generated as $\Sigma = (\sigma_{ij})_{p\times p}$ with $\sigma_{ij} = 0.5^{|i-j|}$.

Different examples generate the response Y using various models. Examples 1–3 use three different generalized linear models, Example 4 considers an additive model, Example 5 uses a Poisson regression model from [53], and Example 6 is a logistic regression model. The number of relevant features in these models ranges from 2 to 5. The parameters of the Lasso and F-score are selected from the 5-fold CV. Meanwhile, we conduct two correlation tests, Kendall test and Spearman test, to analyze the relationship between the response and each predictor sample. The p-values associated with the relevant predictors, in relation to the response, are consistently lower than 0.05, and a notable proportion of them are below 10^{-6}. These results demonstrate strong and robust correlations between the relevant predictors and the response. This observation aligns well with the context of the underlying dataset. The performance is evaluated using true positive rate (TPR), true negative rate (TNR), balanced accuracy, and the number of selected features abbreviated as model size, which are common in feature selection, i.e., [29, 30, 34, 53]. We implement the program using R code and conduct the computational analysis on a standard laptop computer with a 2.30 GHz Intel Core i7-11800H processor. Tables 1–6 summarize the average results of Examples 1–6 based on 200 simulations, respectively. Due to the limited computing power and the R software limitation, the entries of RFE among $p = 2000$ are missing. The sample standard deviation is shown in parentheses. The best results under each dimension are highlighted in all the tables.

As shown in Tables 1–6, the performance of the FK-RFE archives high balanced accuracy and small model size compared to other methods across all examples, while consistently achieving a high TPR and TNR. Tables 1–6 show that FK-RFE performs significantly better than FBED in terms of TPR, suggesting that the latter is less effective in selecting relevant features. In particular, as shown in Table 5, FK-RFE consistently achieves better balanced accuracy, TPR, and TNR than other methods.

In Examples 1–4, FK-RFE has a slightly lower TPR than some other methods, with a difference of approximately 0.15. However, this is because other methods tend to select a larger number of features, including both relevant and irrelevant ones, while FK-RFE strikes a better balance between them. In Example 6, FK-RFE slightly outperforms RFE and significantly outperforms F-score technique for binary dataset. Moreover, as the dimension increases and the number of irrelevant features grows, FK-RFE's selection accuracy remains stable. In summary, across various models and dimensions, FK-RFE consistently achieves high selection accuracy and outperforms other methods in terms of balanced accuracy and model size.

5. Applications

In this section, we demonstrate the effectiveness of the proposed method using the Tecator dataset and Daily Demand Orders dataset. We compare the proposed method with RFE, FRS, and Lasso. However, it should be noted that FBED is not applicable for prediction as it is specifically designed for feature selection. For both datasets, we increase the dimensionality by adding noise. Both datasets are continuous and high dimensional for they have small sample sizes and large numbers of features.

5.1. Tecator Data. The first real example is to analyze the Tecator dataset, which was collected using the near infrared transmission (NIT) principle by the Tecator infratec food and feed analyzer within the wavelength range of 850–1050 nm. The dataset consists of 240 samples and 100 predictors, representing absorbance channel spectra, with the response being the proportion of fat in finely chopped meat. The dataset was previously analyzed by Mai and Zou [53] and can be accessed at https://lib.stat.cmu.edu/datasets/tecator.

We randomly select 200 samples as the training set and use the remaining 40 samples as the testing set. In addition to the 100 predictors in the original dataset, we add 900 independent noise variables following the standard normal distribution and simulate 50 times. We evaluated the effectiveness of the methods in terms of model selection performance, fitting

Table 1: Performance comparison under Example 1.

Methods	p	Balanced accuracy	Model size	TPR	TNR
FK-RFE	100	**0.848 (0.12)**	9.50 (5.78)	0.757 (0.26)	0.940 (0.05)
	300	0.868 (0.12)	9.36 (5.66)	0.755 (0.25)	0.981 (0.02)
	500	0.855 (0.12)	9.02 (5.52)	0.721 (0.25)	**0.989 (0.01)**
	2000	0.867 (0.13)	**9.89 (5.73)**	0.737 (0.26)	**0.997 (0.01)**
RFE	100	0.696 (0.15)	48.24 (36.63)	0.854 (0.23)	0.537 (0.38)
	300	0.660 (0.13)	186.98 (94.46)	0.938 (0.14)	0.382 (0.32)
	500	0.689 (0.13)	286.24 (141.96)	**0.947 (0.13)**	0.431 (0.29)
	2000	—	—	—	—
FRS	100	0.699 (0.14)	50.04 (35.42)	**0.878 (0.22)**	0.519 (0.37)
	300	**0.873 (0.06)**	62.66 (28.54)	**0.942 (0.15)**	0.804 (0.10)
	500	**0.903 (0.07)**	66.70 (28.40)	0.932 (0.16)	0.875 (0.06)
	2000	**0.922 (0.07)**	72.57 (25.13)	**0.878 (0.15)**	0.966 (0.01)
Lasso	100	0.613 (0.10)	62.18 (27.01)	0.836 (0.17)	0.789 (0.28)
	300	0.723 (0.10)	79.72 (24.15)	0.704 (0.19)	0.742 (0.08)
	500	0.730 (0.11)	57.88 (34.20)	0.572 (0.24)	0.889 (0.07)
	2000	0.725 (0.11)	87.52 (18.34)	0.492 (0.22)	0.957 (0.01)
FBED	100	0.642 (0.07)	**3.40 (1.02)**	0.304 (0.13)	**0.980 (0.01)**
	300	0.637 (0.06)	**5.59 (1.25)**	0.288 (0.12)	**0.986 (0.01)**
	500	0.621 (0.06)	**6.99 (1.30)**	0.254 (0.12)	0.988 (0.01)
	2000	0.591 (0.06)	11.40 (1.29)	0.187 (0.12)	0.995 (0.01)

The best results under each dimension are highlighted in all the tables.

Table 2: Performance comparison under Example 2.

Methods	p	Balanced accuracy	Model size	TPR	TNR
FK-RFE	100	**0.805 (0.19)**	7.31 (5.54)	0.670 (0.38)	0.939 (0.06)
	300	0.816 (0.18)	7.51 (5.39)	0.653 (0.37)	0.979 (0.02)
	500	0.755 (0.19)	7.30 (5.55)	0.523 (0.38)	0.987 (0.01)
	2000	0.712 (0.17)	**7.36 (5.40)**	0.428 (0.35)	**0.997 (0.01)**
RFE	100	0.658 (0.18)	42.81 (38.48)	0.738 (0.36)	0.578 (0.39)
	300	0.567 (0.09)	251.34 (58.62)	**0.970 (0.15)**	0.163 (0.20)
	500	0.589 (0.11)	377.86 (106.34)	**0.933 (0.20)**	0.245 (0.21)
	2000	—	—	—	—
FRS	100	0.630 (0.16)	49.49 (41.02)	0.750 (0.37)	0.510 (0.41)
	300	0.767 (0.14)	62.98 (33.86)	0.740 (0.34)	0.794 (0.11)
	500	0.771 (0.16)	71.76 (29.40)	0.683 (0.34)	0.859 (0.06)
	2000	0.701 (0.19)	78.05 (25.96)	0.440 (0.38)	0.961 (0.01)
Lasso	100	0.679 (0.14)	62.62 (27.60)	**0.978 (0.12)**	0.781 (0.28)
	300	**0.844 (0.11)**	68.91 (27.82)	0.913 (0.20)	0.775 (0.09)
	500	**0.850 (0.11)**	88.44 (17.34)	0.875 (0.22)	0.826 (0.03)
	2000	**0.750 (0.17)**	95.29 (8.60)	0.548 (0.33)	0.953 (0.01)
FBED	100	0.747 (0.15)	**3.11 (1.07)**	0.515 (0.30)	**0.979 (0.01)**
	300	0.675 (0.15)	**5.47 (1.21)**	0.365 (0.29)	**0.984 (0.01)**
	500	0.683 (0.16)	**6.83 (1.24)**	0.378 (0.33)	**0.988 (0.01)**
	2000	0.621 (0.15)	11.55 (1.21)	0.248 (0.30)	0.994 (0.01)

The best results under each dimension are highlighted in all the tables.

performance, and prediction performance. Model size and wrong selection (the number of selections from generated noises) are used to measure model selection performance, and the results are presented in Table 7 (the best results are highlighted). We also calculate mean square error,

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2, \tag{16}$$

mean absolute error,

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|, \tag{17}$$

and mean absolute percentage error,

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|\hat{y}_i - y_i\right|}{y_i} \times 100\%, \tag{18}$$

Table 3: Performance comparison under Example 3.

Methods	p	Balanced accuracy	Model size	TPR	TNR
FK-RFE	100	**0.815** (0.13)	7.72 (5.40)	0.695 (0.28)	0.935 (0.05)
	300	0.818 (0.13)	7.64 (5.39)	0.658 (0.28)	0.979 (0.02)
	500	0.826 (0.14)	7.60 (5.38)	0.665 (0.29)	0.987 (0.01)
	2000	0.822 (0.15)	7.13 (5.57)	0.648 (0.30)	0.997 (0.01)
RFE	100	0.710 (0.16)	42.01 (36.95)	0.833 (0.25)	0.588 (0.37)
	300	0.615 (0.14)	208.80 (94.32)	**0.925 (0.18)**	0.306 (0.32)
	500	0.591 (0.15)	360.16 (114.44)	**0.903 (0.20)**	0.280 (0.23)
	2000	—	—	—	—
FRS	100	0.812 (0.19)	26.19 (35.75)	**0.873 (0.22)**	0.751 (0.36)
	300	**0.869 (0.12)**	40.88 (38.53)	0.870 (0.23)	0.869 (0.13)
	500	**0.894 (0.11)**	51.78 (36.58)	0.888 (0.23)	0.900 (0.07)
	2000	**0.923 (0.11)**	46.56 (39.69)	**0.868 (0.23)**	0.978 (0.02)
Lasso	100	0.656 (0.17)	42.46 (41.46)	0.730 (0.34)	0.582 (0.42)
	300	0.694 (0.15)	30.80 (38.71)	0.488 (0.34)	0.900 (0.13)
	500	0.643 (0.15)	15.05 (30.59)	0.315 (0.33)	0.971 (0.06)
	2000	0.801 (0.16)	40.97 (38.95)	0.623 (0.32)	0.980 (0.02)
FBED	100	0.814 (0.11)	**2.44 (1.17)**	0.640 (0.23)	**0.988 (0.01)**
	300	0.791 (0.11)	**3.41 (1.54)**	0.590 (0.21)	**0.993 (0.01)**
	500	0.775 (0.11)	**3.85 (1.56)**	0.555 (0.23)	**0.995 (0.01)**
	2000	0.753 (0.12)	**4.84 (1.51)**	0.508 (0.23)	**0.998 (0.01)**

The best results under each dimension are highlighted in all the tables.

Table 4: Performance comparison under Example 4.

Methods	p	Balanced accuracy	Model size	TPR	TNR
FK-RFE	100	**0.801 (0.14)**	10.50 (6.19)	0.683 (0.29)	0.919 (0.06)
	300	**0.804 (0.16)**	10.44 (6.03)	0.634 (0.32)	0.973 (0.02)
	500	0.813 (0.15)	11.47 (6.35)	0.644 (0.31)	0.982 (0.01)
	2000	0.765 (0.14)	11.52 (6.12)	0.535 (0.29)	**0.995 (0.01)**
RFE	100	0.643 (0.16)	49.38 (34.36)	0.768 (0.34)	0.518 (0.35)
	300	0.621 (0.20)	188.36 (96.48)	**0.868 (0.22)**	0.375 (0.33)
	500	0.625 (0.21)	279.77 (152.97)	0.808 (0.27)	0.442 (0.31)
	2000	—	—	—	—
FRS	100	0.640 (0.16)	56.76 (34.93)	**0.836 (0.26)**	0.544 (0.36)
	300	0.804 (0.14)	63.14 (30.11)	0.811 (0.29)	0.798 (0.10)
	500	**0.859 (0.12)**	69.07 (27.76)	**0.850 (0.25)**	0.868 (0.06)
	2000	**0.919 (0.11)**	69.81 (28.30)	**0.871 (0.22)**	0.967 (0.01)
Lasso	100	0.637 (0.15)	48.11 (23.45)	0.745 (0.21)	0.530 (0.24)
	300	0.669 (0.16)	63.98 (25.04)	0.548 (0.27)	0.791 (0.09)
	500	0.652 (0.17)	81.36 (17.67)	0.464 (0.30)	0.840 (0.04)
	2000	0.643 (0.17)	90.10 (10.17)	0.330 (0.34)	0.956 (0.01)
FBED	100	0.604 (0.11)	**3.11 (1.16)**	0.230 (0.22)	**0.977 (0.01)**
	300	0.584 (0.10)	**5.34 (1.26)**	0.184 (0.20)	**0.984 (0.01)**
	500	0.593 (0.11)	**6.75 (1.39)**	0.199 (0.22)	**0.988 (0.01)**
	2000	0.571 (0.10)	**11.32 (1.50)**	0.148 (0.19)	0.995 (0.01)

The best results under each dimension are highlighted in all the tables.

on the training and testing set, respectively. The performance of the RFE and the FRS is based on the random forest, and the results of all approaches on these three metrics are shown in Tables 8 and 9. The sample standard deviation is shown in parentheses.

In Tecator data analysis, FK-RFE consistently outperforms the other methods. As shown in Table 7, the model size and the number of wrong selections of FK-RFE are significantly smaller than those of the other methods. Specifically, FK-RFE reduces the number of wrong selections of RFE by 81%, FRS by 92%, and Lasso by 99%. Furthermore, FK-RFE achieves the lowest MSE, MAE, and MAPE on both the training and testing sets, as shown in Tables 8 and 9. For instance, FK-RFE reduces the predicted MAPE of RFE by 4%, FRS by 13%, and Lasso by 54%.

5.2. Daily Demand Orders Data. The second real example is to analyze daily demand orders data, which was studied by an artificial neural network [82]. The original dataset is a real database of a Brazilian logistics company. It was collected

Table 5: Performance comparison under Example 5.

Methods	p	Balanced accuracy	Model size	TPR	TNR
FK-RFE	100	**0.977 (0.02)**	6.43 (3.27)	**1.000 (0.00)**	0.955 (0.03)
	300	**0.991 (0.02)**	6.36 (3.04)	**0.998 (0.04)**	**0.985 (0.01)**
	500	**0.994 (0.02)**	**6.44 (3.05)**	**0.998 (0.04)**	**0.991 (0.01)**
	2000	**0.998 (0.02)**	**6.56 (3.33)**	**0.998 (0.04)**	**0.998 (0.01)**
RFE	100	0.945 (0.05)	11.13 (6.01)	0.983 (0.09)	0.906 (0.06)
	300	0.982 (0.01)	12.80 (7.52)	1.000 (0.00)	0.964 (0.03)
	500	0.978 (0.05)	15.55 (8.41)	0.983 (0.09)	0.973 (0.02)
	2000	—	—	—	—
FRS	100	0.745 (0.18)	42.00 (34.81)	0.900 (0.20)	0.590 (0.35)
	300	0.824 (0.13)	61.63 (36.33)	0.850 (0.27)	0.799 (0.12)
	500	0.841 (0.14)	52.03 (36.94)	0.783 (0.31)	0.899 (0.07)
	2000	0.953 (0.08)	57.77 (35.96)	0.933 (0.17)	0.972 (0.02)
Lasso	100	0.585 (0.15)	82.38 (28.60)	0.990 (0.07)	0.180 (0.29)
	300	0.563 (0.11)	22.50 (39.55)	0.200 (0.30)	0.926 (0.13)
	500	0.724 (0.15)	71.62 (39.04)	0.590 (0.30)	0.859 (0.08)
	2000	0.592 (0.13)	33.84 (46.46)	0.200 (0.29)	0.983 (0.02)
FBED	100	0.747 (0.08)	**3.20 (1.16)**	0.517 (0.16)	**0.978 (0.01)**
	300	0.743 (0.11)	**5.43 (1.63)**	0.500 (0.23)	0.985 (0.01)
	500	0.753 (0.10)	6.60 (1.67)	0.517 (0.21)	0.989 (0.01)
	2000	0.665 (0.15)	8.23 (1.96)	0.333 (0.30)	0.996 (0.01)

The best results under each dimension are highlighted in all the tables.

Table 6: Performance comparison under Example 6.

Methods	p	Balanced accuracy	Model size	TPR	TNR
FK-RFE	100	**0.870 (0.11)**	9.72 (5.13)	**0.807 (0.21)**	0.932 (0.05)
	300	**0.839 (0.10)**	16.92 (9.92)	**0.725 (0.20)**	0.953 (0.03)
	500	**0.825 (0.09)**	19.65 (10.65)	**0.684 (0.19)**	0.966 (0.02)
	2000	0.779 (0.11)	31.59 (17.56)	0.573 (0.23)	0.985 (0.01)
RFE	100	**0.870 (0.10)**	9.77 (5.68)	**0.807 (0.19)**	0.932 (0.05)
	300	0.825 (0.12)	16.52 (10.99)	0.697 (0.25)	0.953 (0.03)
	500	0.810 (0.12)	20.14 (12.04)	0.655 (0.24)	0.964 (0.02)
	2000	—	—	—	—
FRS	100	0.819 (0.08)	14.74 (8.36)	0.76 (0.16)	0.878 (0.08)
	300	0.809 (0.08)	25.08 (15.17)	0.693 (0.18)	0.925 (0.05)
	500	0.801 (0.07)	31.33 (18.96)	0.660 (0.16)	0.942 (0.03)
	2000	**0.793 (0.07)**	40.89 (20.85)	**0.605 (0.14)**	0.981 (0.01)
Lasso	100	0.729 (0.05)	11.45 (2.68)	0.554 (0.11)	0.904 (0.02)
	300	0.740 (0.04)	11.38 (3.06)	0.512 (0.07)	0.968 (0.01)
	500	0.743 (0.03)	11.13 (3.19)	0.504 (0.06)	0.982 (0.01)
	2000	0.736 (0.04)	22.66 (4.51)	0.482 (0.08)	0.990 (0.01)
FBED	100	0.717 (0.05)	**3.24 (0.92)**	0.450 (0.10)	**0.985 (0.01)**
	300	0.699 (0.06)	**4.95 (1.16)**	0.409 (0.12)	**0.989 (0.01)**
	500	0.703 (0.06)	**6.18 (1.11)**	0.415 (0.12)	**0.991 (0.01)**
	2000	0.670 (0.06)	**8.18 (1.14)**	0.344 (0.12)	**0.997 (0.01)**
F-score	100	0.773 (0.07)	20 (0)	0.725 (0.14)	0.822 (0.01)
	300	0.770 (0.06)	17 (0)	0.605 (0.13)	0.951 (0.01)
	500	0.787 (0.06)	14 (0)	0.598 (0.12)	0.977 (0.01)
	2000	0.767 (0.05)	18 (0)	0.542 (0.11)	0.992 (0.01)

The best results under each dimension are highlighted in all the tables.

Table 7: The performance of model selection.

	FK-RFE	RFE	FRS	Lasso
Model size	**20.14 (15.80)**	21.76 (10.47)	93.41 (6.19)	163.27 (7.84)
Wrong selection	**0.15 (0.45)**	0.80 (1.03)	1.90 (1.86)	162.86 (7.66)

The best results under each dimension are highlighted in all the tables.

Table 8: The performance of fitting error.

	FK-RFE	RFE	FRS	Lasso
MSE	**0.049 (0.004)**	0.049 (0.004)	0.058 (0.005)	0.081 (0.017)
MAE	**0.151 (0.006)**	0.154 (0.008)	0.173 (0.009)	0.225 (0.024)
MAPE (%)	**48.60 (6.748)**	49.32 (7.433)	51.80 (5.873)	52.94 (8.150)

The best results under each dimension are highlighted in all the tables.

Table 9: The performance of prediction error.

	FK-RFE	RFE	FRS	Lasso
MSE	**0.255 (0.103)**	0.283 (0.095)	0.342 (0.116)	1.672 (0.306)
MAE	**0.354 (0.067)**	0.380 (0.065)	0.432 (0.069)	1.026 (0.100)
MAPE (%)	**108.18 (54.5)**	113.62 (55.031)	125.65 (52.64)	233.77 (117.301)

The best results under each dimension are highlighted in all the tables.

Table 10: The performance of model selection.

	FK-RFE	RFE	FRS	Lasso
Model size	**5.96 (3.12)**	11.18 (9.64)	6.22 (4.55)	6.06 (2.59)
Wrong selection	**1.42 (2.27)**	5.22 (8.92)	1.66 (3.16)	5.46 (2.29)

The best results under each dimension are highlighted in all the tables.

Table 11: The performance of fitting error.

	FK-RFE	RFE	FRS	Lasso
MSE	0.038 (0.008)	**0.036 (0.008)**	0.045 (0.016)	0.664 (0.057)
MAE	0.113 (0.011)	**0.110 (0.011)**	0.125 (0.027)	0.620 (0.030)
MAPE (%)	45.535 (15.195)	**37.524 (9.153)**	43.712 (14.330)	111.888 (15.552)

The best results under each dimension are highlighted in all the tables.

Table 12: The performance of prediction error.

	FK-RFE	RFE	FRS	Lasso
MSE	**0.189 (0.125)**	0.256 (0.222)	0.221 (0.210)	0.994 (0.413)
MAE	**0.274 (0.073)**	0.308 (0.112)	0.279 (0.125)	0.737 (0.134)
MAPE (%)	**94.421 (54.741)**	96.514 (53.096)	85.201 (43.064)	139.322 (37.936)

The best results under each dimension are highlighted in all the tables.

during 60 days, which has 12 predictive attributes and a target that is the total of orders for daily treatment and can be accessed at https://archive.ics.uci.edu/ml/datasets/Daily+Demand+Forecasting+Orders.

We randomly split 60 samples into a training set (40 samples) and a testing set (20 samples) with 50 times simulations. We add 900 independent noise variables and use the same performance measures as the above real data. The experimental results are shown in Tables 10–12, in which the best results are highlighted.

In this data analysis, FK-RFE consistently outperforms other methods in terms of model selection and prediction. Table 10 demonstrates that FK-RFE achieves the smallest model size and wrong selection compared to the other methods. In addition, Table 12 shows that FK-RFE leads to reduced MSE, MAE, and MAPE values compared to other methods. For example, FK-RFE reduces the predicted MSE of RFE by 26%, FRS by 14%, and Lasso by 80%. In terms of fitting, as shown in Table 11, FK-RFE performs comparably to RFE and achieves the best results.

6. Summary

In this paper, we introduce a novel feature selection procedure combining the filter and wrapper technique, named FK-RFE. This method is designed to efficiently handle complex ultrahigh dimensional datasets without being limited by model assumptions. We demonstrate that the proposed method is selection consistent and L_2 consistent under mild conditions. We evaluate the performance of the proposed method under various types of data. Results obtained from simulations and applications show that FK-RFE outperforms some existing methods, highlighting its superior efficiency in feature selection. Overall, FK-RFE is fast, accurate, and model-free, making it a useful and efficient technique for resolving feature selection issues.

On the other hand, there exist some limitations to the proposed method that deserve further improvement. Due to the limitation of the fused Kolmogorov filter and the random forest, FK-RFE does not consider the capacity of model learning in the first filter phase. Furthermore, FK-RFE also has limitations, particularly when dealing with specific types of datasets, such as unbalanced datasets. In such cases, for applying random forest, FK-RFE is unsuitable. It would be interesting to explore novel techniques to address the challenges posed by these kinds of data.

Exploring the potential of combining FK-RFE with other feature selection methods, such as information gain or correlation-based methods, could be a promising direction for future research. Moreover, investigating the interpretability of the selected features by FK-RFE and the potential for discovering novel biomarkers or causal relationships in complex systems could be a fruitful area for further investigation. Applying FK-RFE to real-world problems in fields such as bioinformatics, finance, or image analysis could provide valuable insights into its practical applications and limitations.

Appendix

Proof 1. As shown in Algorithm 1, let $V_0, V_1, \ldots, V_{d_n-1}$ be the sequence of active sets selected during iterations, in which V_0 is selected by the filter phase and $V_1, \ldots, V_{d_n-1}$ are obtained by eliminating one variable at each step in the wrapper phase. By the nature of the sequential procedure, this is a nested sequence, i.e.,

$$V_0 \supset V_1 \supset V_2 \supset \ldots \supset V_{d_n-1}. \tag{A.1}$$

We aim to prove the convergence of the algorithm. It suffices to show that there exists a step $k \in \{0, 1, 2 \ldots, d_n - 1\}$ such that the performance error of the random forest estimation under the model V_k is the minimum and V_k is the true model S.

As mentioned in Theorem 1 of [53], under conditions C1 and C2, we have $\{S \subset V_0\}$. Under the assumption of the importance measure,

$$\lim_{M \longrightarrow \infty} E\left(\widehat{I}\left(X_j\right)\right) = I\left(X_j\right). \tag{A.2}$$

We have the empirical permutation importance measure is unbiased. Based on the definition of the permutation importance measure, for $j \in S^c$,

$$\begin{aligned} I\left(X_j\right) &= E\left[\left(Y - f\left(X_{(j)}\right)\right)^2\right] - E\left[(Y - f(X))^2\right] \\ &= E\left[(Y - f(X))^2\right] - E\left[(Y - f(X))^2\right] = 0. \end{aligned} \tag{A.3}$$

On the other hand, for $j \in S$,

$$\begin{aligned} I\left(X_j\right) &= E\left[\left(Y - f\left(X_{(j)}\right)\right)^2\right] - E\left[(Y - f(X))^2\right] \\ &= E\left[\left((Y - f(X)) + \left(f(X) - f\left(X_{(j)}\right)\right)\right)^2\right] - E\left[(Y - f(X))^2\right] \\ &= E\left[\left(f(X) - f\left(X_{(j)}\right)\right)^2\right] + E\left[\epsilon\left(f(X) - f\left(X_{(j)}\right)\right)\right] \\ &= E\left[\left(f(X) - f\left(X_{(j)}\right)\right)^2\right]. \end{aligned} \tag{A.4}$$

The last equality follows from the assumption that ϵ is independent of X and $X_{(j)}$. This leads to $E[\epsilon(f(X)] = 0$ and $E[\epsilon(f(X_{(j)})] = 0$. Thus, we can obtain that.

$$\begin{aligned} &I\left(X_j\right) = 0, \quad \text{for } j \in S^c, \\ &I\left(X_j\right) > 0, \quad \text{for } j \in S. \end{aligned} \tag{A.5}$$

Noted that at each iteration, the wrapper phase eliminates the least important variable with the smallest value of the permutation importance measure. Based on the abovementioned result (A.5), we have that the unimportant variables would be eliminated first. Thus, set $k = d_n - q$. We have that there exists an active subset that $V_k = S$. Under the requirement of the random forest estimator, we have

$$\lim_{n \longrightarrow \infty} E[\overline{f} - f]^2 = 0. \tag{A.6}$$

It means that the random forest under model V_k has the best performance and thus V_k can be selected as the optimal model according to the criterion, i.e., $\widehat{S} = V_k$, completing the proof. □

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Disclosure

A preprint has previously been published [83].

Conflicts of Interest

The authors declare that there are no potential conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 12001557); the Fundamental Research Funds for the Central University (Grant no. PHD2023-054); and the Emerging Interdisciplinary Project, Program for Innovation Research, and the Disciplinary Funds in Central University of Finance and Economics.

References

[1] A. Ghaemi, E. Rashedi, A. M. Pourrahimi, M. Kamandar, and F. Rahdari, "Automatic channel selection in EEG signals for classification of left or right hand movement in Brain Computer Interfaces using improved binary gravitation search algorithm," *Biomedical Signal Processing and Control*, vol. 33, pp. 109–118, 2017.

[2] K. Benidis, Y. Feng, and D. P. Palomar, "Sparse portfolios for high-dimensional financial index tracking," *IEEE Transactions on Signal Processing*, vol. 66, no. 1, pp. 155–170, 2018.

[3] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: a survey," *Information*, vol. 10, no. 4, p. 150, 2019.

[4] O. A. Alomari, S. N. Makhadmeh, M. A. Al-Betar et al., "Gene selection for microarray data classification based on Gray Wolf Optimizer enhanced with TRIZ-inspired operators," *Knowledge-Based Systems*, vol. 223, Article ID 107034, 2021.

[5] M. A. Awadallah, A. I. Hammouri, M. A. Al-Betar, M. S. Braik, and M. A. Elaziz, "Binary Horse herd optimization algorithm with crossover operators for feature selection," *Computers in Biology and Medicine*, vol. 141, Article ID 105152, 2022.

[6] M. A. Awadallah, M. A. Al-Betar, M. S. Braik, A. I. Hammouri, I. A. Doush, and R. A. Zitar, "An enhanced binary Rat Swarm Optimizer based on local-best concepts of PSO and collaborative crossover operators for feature selection," *Computers in Biology and Medicine*, vol. 147, Article ID 105675, 2022.

[7] S. Xia, Y. Yang, and H. Yang, "High-dimensional sparse portfolio selection with nonnegative constraint," *Applied Mathematics and Computation*, vol. 443, Article ID 127766, 2023.

[8] A. K. Jain, P. Duin, and J. Mao, "Statistical pattern recognition: a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[9] R. Sheikhpour, M. A. Sarram, S. Gharaghani, and M. A. Z. Chahooki, "A survey on semi-supervised feature selection methods," *Pattern Recognition*, vol. 64, pp. 141–158, 2017.

[10] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: a review," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 4, pp. 1060–1073, 2022.

[11] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[12] M. Miri, M. B. Dowlatshahi, A. Hashemi, M. K. Rafsanjani, B. B. Gupta, and W. Alhalabi, "Ensemble feature selection for multi-label text classification: an intelligent order statistics approach," *International Journal of Intelligent Systems*, vol. 37, no. 12, pp. 11319–11341, 2022.

[13] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[14] J. H. Cho and P. U. Kurup, "Decision tree approach for classification and dimensionality reduction of electronic nose data," *Sensors and Actuators B: Chemical*, vol. 160, no. 1, pp. 542–548, 2011.

[15] J. Q. Fan and R. Z. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[16] C. H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Annals of Statistics*, vol. 38, no. 2, pp. 894–942, 2010.

[17] X. Y. Liu, S. B. Wu, W. Q. Zeng, Z. J. Yuan, and H. B. Xu, "LogSum + L_2 penalized logistic regression model for biomarker selection and cancer classification," *Scientific Reports*, vol. 10, no. 1, Article ID 22125, 2020.

[18] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, no. 2, pp. 407–451, 2004.

[19] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, "Pathwise coordinate optimization," *Annals of Applied Statistics*, vol. 1, no. 2, pp. 302–332, 2007.

[20] J. Friedman, T. Hastie, and R. J. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.

[21] M. Yuan and Y. Xu, "Feature screening strategy for non-convex sparse logistic regression with log sum penalty," *Information Sciences*, vol. 624, pp. 732–747, 2023.

[22] P. Zhao and B. Yu, "On model selection consistency of lasso," *Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.

[23] A. Janecek, W. Gansterer, M. Demel, and G. Ecker, "On the relationship between feature selection and classification accuracy," in *Proceedings of the 2008 International Conference on New Challenges for Feature Selection in Data Mining and Knowledge Discovery*, pp. 90–105, PMLR, Antwerp, Belgium, September 2008.

[24] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[25] K. Miyahara and M. J. Pazzani, "Collaborative filtering with the simple bayesian classifier," in *Proceedings of the Pacific Rim International conference on artificial intelligence*, pp. 679–689, Springer, Berlin, Germany, August 2000.

[26] M. Bennasar, Y. Hicks, and R. Setchi, "Feature selection using joint mutual information maximisation," *Expert Systems with Applications*, vol. 42, no. 22, pp. 8520–8532, 2015.

[27] S. Ramírez-Gallego, I. Lastra, D. Martínez-Rego et al., "Fast-mRMR: fast minimum redundancy maximum relevance algorithm for high-dimensional big data," *International Journal of Intelligent Systems*, vol. 32, no. 2, pp. 134–152, 2017.

[28] Y. W. Chen and C. J. Lin, "Combining SVMs with various feature selection strategies," *Feature extraction: Foundations and Applications*, pp. 315–324, 2006.

[29] N. Q. K. Le and Y. Y. Ou, "Prediction of FAD binding sites in electron transport proteins according to efficient radial basis function networks and significant amino acid pairs," *BMC Bioinformatics*, vol. 17, no. 1, pp. 298–313, 2016.

[30] N. Q. K. Le, T. T. Huynh, E. K. Y. Yapp, and H. Y. Yeh, "Identification of clathrin proteins by incorporating hyper-parameter optimization in deep learning and PSSM profiles," *Computer Methods and Programs in Biomedicine*, vol. 177, pp. 81–88, 2019.

[31] J. Pirgazi, M. Alimoradi, T. Esmaeili Abharian, and M. H. Olyaee, "An Efficient hybrid filter-wrapper metaheuristic-based gene selection method for high dimensional datasets," *Scientific Reports*, vol. 9, no. 1, Article ID 18580, 2019.

[32] H. Zhou, Y. Zhang, Y. Zhang, and H. Liu, "Feature selection based on conditional mutual information: minimum

conditional relevance and minimum conditional redundancy," *Applied Intelligence*, vol. 49, no. 3, pp. 883–896, 2019.

[33] L. Wang, S. Jiang, and S. Jiang, "A feature selection method via analysis of relevance, redundancy, and interaction," *Expert Systems with Applications*, vol. 183, Article ID 115365, 2021.

[34] J. Q. Fan and J. C. Lv, "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society-Series B: Statistical Methodology*, vol. 70, no. 5, pp. 849–911, 2008.

[35] J. Fan, R. Li, C. H. Zhang, and H. Zou, *Statistical Foundations of Data Science*, CRC Press, Boca Raton, FL, USA, 2020.

[36] W. Liu and R. Li, "Variable selection and feature screening," in *Macroeconomic Forecasting in the Era of Big Data: Theory and Practice*, pp. 293–326, Springer, Berlin, Germany, 2020.

[37] J. Fan and R. Song, "Sure independence screening in generalized linear models with NP-dimensionality," *Annals of Statistics*, vol. 38, no. 6, pp. 3567–3604, 2010.

[38] S. D. Zhao and Y. Li, "Principled sure independence screening for Cox models with ultra-high-dimensional covariates," *Journal of Multivariate Analysis*, vol. 105, no. 1, pp. 397–411, 2012.

[39] P. Xu, L. Zhu, and Y. Li, "Ultrahigh dimensional time course feature selection," *Biometrics*, vol. 70, no. 2, pp. 356–365, 2014.

[40] J. Fan, Y. Feng, and R. Song, "Nonparametric independence screening in sparse ultra-high-dimensional additive models," *Journal of the American Statistical Association*, vol. 106, no. 494, pp. 544–557, 2011.

[41] M. Y. Cheng, T. Honda, J. Li, and H. Peng, "Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data," *Annals of Statistics*, vol. 42, no. 5, pp. 1819–1849, 2014.

[42] M. Y. Cheng, T. Honda, and J. T. Zhang, "Forward variable selection for sparse ultra-high dimensional varying coefficient models," *Journal of the American Statistical Association*, vol. 111, no. 515, pp. 1209–1221, 2016.

[43] W. Chu, R. Li, and M. Reimherr, "Feature screening for time-varying coefficient models with ultrahigh dimensional longitudinal data," *Annals of Applied Statistics*, vol. 10, no. 2, pp. 596–617, 2016.

[44] W. Chu, R. Li, J. Liu, and M. Reimherr, "Feature selection for generalized varying coefficient mixed-effect models with application to obesity GWAS," *Annals of Applied Statistics*, vol. 14, no. 1, pp. 276–298, 2020.

[45] F. G. Blanchet, P. Legendre, and D. Borcard, "Forward selection of explanatory variables," *Ecology*, vol. 89, no. 9, pp. 2623–2632, 2008.

[46] J. A. Fernández Pierna, O. Abbas, V. Baeten, and P. Dardenne, "A backward variable selection method for PLS regression (BVSPLS)," *Analytica Chimica Acta*, vol. 642, no. 1-2, pp. 89–93, 2009.

[47] S. Maldonado, R. Weber, and F. Famili, "Feature selection for high-dimensional class-imbalanced data sets using support vector machines," *Information Sciences*, vol. 286, pp. 228–246, 2014.

[48] G. Borboudakis and I. Tsamardinos, "Forward-backward selection with early dropping," *Journal of Machine Learning Research*, vol. 20, no. 1, pp. 276–314, 2019.

[49] Q. Zheng, H. G. Hong, and Y. Li, "Building generalized linear models with ultrahigh dimensional features: a sequentially conditional approach," *Biometrics*, vol. 76, no. 1, pp. 47–60, 2020.

[50] T. Honda and C. T. Lin, "Forward variable selection for sparse ultra-high-dimensional generalized varying coefficient models," *Japanese Journal of Statistics and Data Science*, vol. 4, no. 1, pp. 151–179, 2021.

[51] L. P. Zhu, L. Li, R. Li, and L. X. Zhu, "Model-free feature screening for ultrahigh-dimensional data," *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1464–1475, 2011.

[52] X. He, L. Wang, and H. G. Hong, "Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data," *Annals of Statistics*, vol. 41, no. 1, pp. 342–369, 2013.

[53] Q. Mai and H. Zou, "The fused Kolmogorov filter: a nonparametric model-free screening method," *Annals of Statistics*, vol. 43, no. 4, pp. 1471–1497, 2015.

[54] W. Liu, Y. Ke, J. Liu, and R. Li, "Model-free feature screening and fdr control with knockoff features," *Journal of the American Statistical Association*, vol. 117, no. 537, pp. 428–443, 2022.

[55] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1/3, pp. 389–422, 2002.

[56] X. Lin, F. Yang, L. Zhou et al., "A support vector machine-recursive feature elimination feature selection method based on artificial contrast variables and mutual information," *Journal of Chromatography B*, vol. 910, pp. 149–155, 2012.

[57] Y. Guo, Z. Zhang, and F. Tang, "Feature selection with kernelized multi-class support vector machine," *Pattern Recognition*, vol. 117, Article ID 107988, 2021.

[58] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Correlation and variable importance in random forests," *Statistics and Computing*, vol. 27, no. 3, pp. 659–678, 2017.

[59] S. Rumao, *Exploration of Variable Importance and Variable Selection Techniques in Presence of Correlated Variables*, Rochester Institute of Technology, Rochester, NY, USA, 2019.

[60] W. You, Z. Yang, and G. Ji, "PLS-based recursive feature elimination for high-dimensional small sample," *Knowledge-Based Systems*, vol. 55, pp. 15–28, 2014.

[61] S. Xia and Y. Yang, "An iterative model-free feature screening procedure: forward recursive selection," *Knowledge-Based Systems*, vol. 246, Article ID 108745, 2022.

[62] B. Seijo-Pardo, I. Porto-Díaz, V. Bolón-Canedo, and A. Alonso-Betanzos, "Ensemble feature selection: homogeneous and heterogeneous approaches," *Knowledge-Based Systems*, vol. 118, pp. 124–139, 2017.

[63] K. B. Duan, J. C. Rajapakse, H. Wang, and F. Azuaje, "Multiple SVM-RFE for gene selection in cancer classification with expression data," *IEEE Transactions on NanoBioscience*, vol. 4, no. 3, pp. 228–234, 2005.

[64] P. A. Mundra and J. C. Rajapakse, "SVM-RFE with MRMR filter for gene selection," *IEEE Transactions on NanoBioscience*, vol. 9, no. 1, pp. 31–37, 2010.

[65] B. Richhariya, M. Tanveer, A. H. Rashid, and A. D. N. Initiative, "Diagnosis of Alzheimer's disease using universum support vector machine based recursive feature elimination (USVM-RFE)," *Biomedical Signal Processing and Control*, vol. 59, Article ID 101903, 2020.

[66] E. M. Senan, M. H. Al-Adhaileh, F. W. Alsaade et al., "Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques," *Journal of Healthcare Engineering*, vol. 2021, no. 10, Article ID 1004767, 10 pages, 2021.

[67] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[68] B. Ramosaj and M. Pauly, "Asymptotic unbiasedness of the permutation importance measure in random forest models," 2019, https://arxiv.org/abs/1912.03306.

[69] V. Svetnik, A. Liaw, C. Tong, and T. Wang, "Application of breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules," in *Classifier Systems: 5th International Workshop, MCS 2004*, pp. 334–343, Springer, Berlin, Germany, 2004.

[70] R. Díaz-Uriarte and S. A. De Andres, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, no. 1, pp. 1–13, 2006.

[71] H. Jeon and S. Oh, "Hybrid-recursive feature elimination for efficient feature selection," *Applied Sciences*, vol. 10, no. 9, p. 3211, 2020.

[72] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[73] H. Wang, "Forward regression for ultra-high dimensional variable screening," *Journal of the American Statistical Association*, vol. 104, no. 488, pp. 1512–1524, 2009.

[74] J. P. Pluim, J. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: a survey," *IEEE Transactions on Medical Imaging*, vol. 22, no. 8, pp. 986–1004, 2003.

[75] W. Liang and Y. Yang, "A sequential stepwise screening procedure for sparse recovery in high-dimensional multi-response models with complex group structures," 2022, http://arxiv.org/abs/2208.06567.

[76] H. Yang and H. L. Liu, "Penalized weighted composite quantile estimators with missing covariates," *Statistical Papers*, vol. 57, no. 1, pp. 69–88, 2016.

[77] C. K. Ing and T. L. Lai, "A stepwise regression method and consistent model selection for high-dimensional sparse linear models," *Statistica Sinica*, vol. 21, no. 4, pp. 1473–1513, 2011.

[78] E. Scornet, G. Biau, and J. P. Vert, "Consistency of random forests," *Annals of Statistics*, vol. 43, no. 4, pp. 1716–1741, 2015.

[79] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010.

[80] M. Denil, D. Matheson, and N. Freitas, "Consistency of online random forests," in *Proceedings of the 30th International Conference on Machine Learning*, pp. 1256–1264, PMLR, Atlanta, Georgia, September 2013.

[81] S. Athey, J. Tibshirani, and S. Wager, "Generalized random forests," *Annals of Statistics*, vol. 47, no. 2, pp. 1148–1178, 2019.

[82] R. P. Ferreira, A. Martiniano, A. Ferreira, A. Ferreira, and R. Sassi, "Study on daily demand forecasting orders using artificial neural network," *IEEE Latin America Transactions*, vol. 14, no. 3, pp. 1519–1525, 2016.

[83] S. Xia and Y. Yang, "A model-free feature selection technique of feature screening and random forest based recursive feature elimination," 2023, https://arxiv.org/abs/2302.07449.