

Article

Feature Importance Ranking Using Interval-Valued Methods and Aggregation Functions for Machine Learning Applications

Aleksander Wojtowicz , Wiesław Paja *  and Urszula Bentkowska 

Faculty of Exact and Technical Sciences, University of Rzeszów, Al. Rejtana 16C, 35-959 Rzeszów, Poland; alwojtowicz@ur.edu.pl (A.W.); ubentkowska@ur.edu.pl (U.B.)

* Correspondence: wpaja@ur.edu.pl

Abstract

Feature selection is one of the key stages in the process of creating machine learning models and conducting data analysis. This paper presents the results of research related to the implementation of a new algorithm for feature selection and ranking based on weighted interval aggregations. It took into account interval importance values obtained from dividing the dataset into subsets. The algorithm was highly effective in identifying relevant features. The results of comparative studies with nine known methods of feature importance assessment are presented. Ten synthetic datasets and five real datasets were used for the experiments. The calculations also included tests of the relevance of the results obtained. In most experiments, the IVWFR algorithm proved to be the best, achieving the best classification results after identifying subsets of relevant features.

Keywords: feature ranking; feature selection; aggregation; interval-valued aggregation; machine learning; classification



Academic Editor: Douglas O'Shaughnessy

Received: 19 October 2025

Revised: 9 November 2025

Accepted: 11 November 2025

Published: 15 November 2025

Citation: Wojtowicz, A.; Paja, W.; Bentkowska, U. Feature Importance Ranking Using Interval-Valued Methods and Aggregation Functions for Machine Learning Applications. *Appl. Sci.* **2025**, *15*, 12130. <https://doi.org/10.3390/app152212130>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Feature selection is one of the key stages in the process of building machine learning models and conducting data analysis. The appropriate selection of a subset of input variables not only allows for dimensionality reduction and improved computational efficiency, but also increases the interpretability of models and reduces the risk of overfitting [1,2]. The literature on this subject describes many approaches to feature selection, including filtering, wrapper, and embedded methods [3,4]. Each of them differs in the mechanism of evaluating the usefulness of individual variables, as well as the trade-off between accuracy and computational costs.

One promising direction for feature selection is methods based on feature relevance aggregation using intervals. Their basic idea is to analyze the stability and repeatability of feature ranks by determining relevance measures in multiple repetitions or data divisions, and then grouping the results in the form of confidence intervals or rank aggregations [5–9]. This makes it possible not only to identify the most important features, but also to estimate the degree of certainty assigned to their position in the ranking.

The interval aggregation approach is used in situations where individual significance measures may be unstable, especially when there are a large number of features and a limited sample size [10]. The use of this type of method increases the resilience of the selection process to random data fluctuations and provides a more reliable picture of the significance of individual variables in the model [11,12].

The aim of this contribution is to carefully examine the method Interval-Valued Weighted Feature Ranking (IVWFR) proposed in [13]. IVWFR is a feature selection method that is built upon the feature importance derived from a classification model used in the method. The proposed algorithm aims to improve the stability and interpretability of feature selection by incorporating uncertainty (represented by the intervals using the idea of interval-valued fuzzy sets) and aggregating information across multiple data folds. In [13], only preliminary results on the new method were provided. Moreover, the synthetic datasets were the only ones applied in the performed experiments. In this contribution, thorough analysis of the IVWFR method and its performance on different datasets is presented. Both synthetic and real datasets are applied in the experiments. Statistical tests are performed to show the significant differences in the classification metric, balanced accuracy, in favor of the new IVWFR method. The results of the IVWFR method are compared with known feature selection methods: RFECV, RFE, ANOVA, Pearson correlation, Spearman correlation, Mutual Information, Random Forest, and XGBoost. In the experiments, RF denotes Random Forest trained on all features, while RF_Importance refers to a two-stage approach: first computing feature importance, then training a new Random Forest on selected features only. Furthermore, the number of features selected by each method is compared on each dataset and the Jaccard similarity index is computed. The results show that the IVWFR method outperforms many of the existing methods. IVWFR may have comparable results with respect to well-established methods like RFECV or RFE; however, its computational complexity is very competitive against time-consuming methods such as RFE or RFECV.

The structure of this paper is as follows: In Section 2, the basic notions connected with interval-valued calculus are recalled. In Section 3, details of the considered method are provided. Section 4 describes the datasets applied in the experiments. Finally, in Sections 5 and 6, the results of the experiments and the discussion are presented.

2. Interval Methods and Aggregation Functions

In the 1970s, Zadeh introduced interval-valued fuzzy sets (IVFSs) to represent ambiguity and uncertainty in practical applications [14]. Since then, IVFSs have been applied in various domains, including decision-making, pattern recognition, and control systems. Compared to conventional fuzzy sets, the use of interval-valued membership degrees allows for a more flexible and comprehensive representation of uncertainty and vagueness. For instance, due to an element's inherent ambiguity or variability, assigning a single membership degree may not be appropriate in certain applications. In such cases, an IVFS can be employed to capture the range of possible values or the degree of uncertainty associated with an element's membership.

One of the key advantages of IVFSs is their ability to handle imprecise and uncertain information more robustly than standard fuzzy sets. In real-world applications, IVFSs provide a more realistic representation of ambiguous and vague information, which can lead to more accurate and reliable decision-making. Furthermore, IVFSs can be effectively used to model complex and multidimensional uncertainty in situations where conventional fuzzy sets may be insufficient [15].

Interval Calculus is a mathematical framework designed for studying and analyzing real number intervals. The set L^I is the set of all subintervals of the unit interval $I = [0, 1]$, i.e., $L^I = \{[\underline{x}, \bar{x}], \underline{x} \leq \bar{x}, \underline{x}, \bar{x} \in [0, 1]\}$, where each interval can be represented by a pair of numbers $[\underline{x}, \bar{x}]$, such that $\underline{x} \leq \bar{x}$.

An interval-valued fuzzy set F in X is defined as a function $F : X \rightarrow L^I$, $F(x) = [\underline{F}(x), \bar{F}(x)] \in L^I$ for $x \in X$, where $\underline{F}(x), \bar{F}(x)$ denote fuzzy sets.

In our experiments, we represent the data with the use of intervals and we need to aggregate intervals. As a result, we use interval-valued aggregations.

Let $\mathbf{x} = [\underline{x}, \bar{x}]$, $\mathbf{y} = [\underline{y}, \bar{y}] \in L^I$. $\mathcal{A} : (L^I)^n \rightarrow L^I$ is called an interval-valued aggregation function (cf. [16]) in L^I if it is increasing, i.e.,

$$\forall_{\mathbf{x}_i, \mathbf{y}_i \in L^I} \mathbf{x}_i \preceq \mathbf{y}_i \Rightarrow \mathcal{A}(\mathbf{x}_1, \dots, \mathbf{x}_n) \preceq \mathcal{A}(\mathbf{y}_1, \dots, \mathbf{y}_n)$$

and $\mathcal{A}(\underbrace{\mathbf{0}, \dots, \mathbf{0}}_{n \times}) = \mathbf{0}$, $\mathcal{A}(\underbrace{\mathbf{1}, \dots, \mathbf{1}}_{n \times}) = \mathbf{1}$, where $\mathbf{0} = [0, 0]$, $\mathbf{1} = [1, 1]$. The classical partial order for intervals is in the form $[\underline{x}, \bar{x}] \preceq [\underline{y}, \bar{y}] \Leftrightarrow \underline{x} \leq \underline{y}$, $\bar{x} \leq \bar{y}$; however, other orders (cf. [17]) or comparability relations for intervals may be used in the above monotonicity condition.

In our experiments, the interval-valued aggregation function based on the weighted arithmetic mean was used

$$\mathcal{A}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = \left[\sum_{i=1}^n \hat{w}_i \cdot \underline{x}_i, \sum_{i=1}^n \hat{w}_i \cdot \bar{x}_i \right] \quad (1)$$

where $\hat{w}_i$ denotes weights and $\sum_{i=1}^n \hat{w}_i = 1$; however, other weighted interval-valued aggregation functions may also be applied.

Below we present the methods of creating weights for aggregating n intervals, representing the importance of a feature F (obtained across n folds) denoted as

$$\mathbf{x}_i = [\underline{x}_i, \bar{x}_i] \quad \text{for } i = 1, 2, \dots, n \quad (2)$$

where $\underline{x}_i$ and $\bar{x}_i$ represent the lower and upper bounds of the i -th interval, respectively. The goal is to combine these intervals into a single interval, where narrower intervals carry greater weight, reflecting their higher certainty. The strategy detailed below was adapted.

3. Algorithm and Experimental Details

3.1. Interval-Valued Weighted Feature Ranking

The Interval-Valued Weighted Feature Ranking (IVWFR) algorithm, proposed in [13], is a feature selection approach based on feature importance measures obtained from a classification model. In this study, the Random Forest classifier was employed for this purpose. The proposed algorithm aims to enhance the stability and interpretability of feature selection by incorporating uncertainty and aggregating importance information across multiple data folds. The IVWFR procedure is summarized in Algorithm 1.

This approach prioritizes robust feature selection and emphasizes the stability of feature importance while mitigating noise across data folds.

IVWFR is especially suitable for high-dimensional datasets. Comparing the computational efficiency of IVWFR to wrapper-based methods, we see that for a dataset with f features, n samples, and k cross-validation folds, the complexity of IVWFR is approximately $\mathcal{O}(k \cdot T_{\text{RF}}(n, f))$, where $T_{\text{RF}}(n, f)$ is the training time of a Random Forest classifier. In contrast, methods like RFE and RFECV perform repeated training on multiple feature subsets, resulting in a typical complexity of $\mathcal{O}(f \cdot T_{\text{RF}}(n, f))$ or higher.

Algorithm 1: IVWFR: Interval-based Weighted Feature Ranking and Selection

X	Feature matrix (n samples, m features)
y	Target vector
k	Number of cross-validation splits
Input: $Classifier$	Base model for importance estimation (e.g., RF)
$aggregation_method$	Strategy to combine importance intervals
$selection_method$	Strategy to select features
α, β	Weights for importance and uncertainty

Output: Selected feature subset
Trained classifier model on selected features

```

1 if  $selection\_method$  is 'contrast' then
2   | Add a random 'contrast' feature to  $X$ ;
3 end
4 Create  $k$  stratified folds from  $(X, y)$ ;
5 foreach fold  $f \in \{1, \dots, k\}$  do
6   | Train  $Classifier_f$  on the training data of fold  $f$ ;
7   | foreach feature  $j \in \{1, \dots, m\}$  do
8     | Compute importance mean  $\mu_{j,f}$  and std  $\sigma_{j,f}$  from  $Classifier_f$ ;
9     |  $I_{j,f} \leftarrow [\max(0, \mu_{j,f} - 1.96\sigma_{j,f}), \mu_{j,f} + 1.96\sigma_{j,f}]$ ;
10    | end
11  end
12 foreach feature  $j \in \{1, \dots, m\}$  do
13   | Aggregate intervals  $\{I_{j,1}, \dots, I_{j,k}\}$  into a single interval  $\bar{I}_j$ ;
14 end
15 Normalize all aggregated intervals  $\bar{I}_j$ ;
16 foreach normalized interval  $\bar{I}_j$  do
17   | Calculate its center  $c_j$  and width  $w_j$ ;
18   | Compute score  $S_j \leftarrow \alpha \cdot c_j + \beta \cdot (1 - w_j)$ ;
19 end
20 Sort features by scores  $S$  in descending order;
21 Select final features  $X_{selected}$  based on the  $selection\_method$ ;
22 if  $selection\_method$  is 'contrast' then
23   | Remove 'contrast' feature from  $X_{selected}$  if present;
24 end
25 Train final  $Classifier$  on  $(X_{selected}, y)$ ;
26 return Features in  $X_{selected}$  and the final  $Classifier$ ;

```

3.2. Cross-Validation and Importance Interval Estimation

The input dataset is divided into several folds using stratified k -fold cross-validation, ensuring that each fold reflects the overall data distribution. The size of each training fold can be adjusted to a desired proportion of the total dataset. For each training fold, a Random Forest classifier is trained. Rather than relying on single-point estimates of feature importance, the algorithm computes confidence intervals for the importance of each feature. These intervals are derived from the distribution of importance values across individual trees within the Random Forest, providing a quantitative measure of the uncertainty associated with each feature's importance.

3.3. Interval-Valued Aggregation, Normalization, and Feature Scoring

The importance intervals for each feature obtained from different cross-validation folds are combined into a single representative interval using interval-valued aggregation functions. This aggregation process employs an interval version of the weighted arithmetic mean, where the weights are determined by the width of each interval and a selected representative point within it (such as the lower bound, upper bound, or midpoint). A small smoothing factor, denoted as ϵ , is incorporated to handle intervals with very small or zero widths. The aggregated intervals are then normalized to a common scale, ensuring comparability across features.

Based on the normalized intervals, a composite score is calculated for each feature. This score reflects a combination of the interval's central tendency (representing the average importance) and its width (representing the uncertainty). The relative contribution of the center and the width in the score calculation can be adjusted, and various types of averaging methods may be employed, such as arithmetic, geometric, or harmonic means. The details of the proposed procedure are summarized below.

Step 1: Calculate the width of each interval

For each interval x_i , the width is calculated as:

$$\text{width}_i = \bar{x}_i - \underline{x}_i \quad (3)$$

Step 2: Determine the representative value for each interval

A representative value r_i is chosen for each interval, which can be one of the following:

- Lower bound: $r_i = \underline{x}_i$;
- Upper bound: $r_i = \bar{x}_i$;
- Midpoint: $r_i = \frac{\underline{x}_i + \bar{x}_i}{2}$.

The midpoint is used as the default representative value in the presented method.

Step 3: Calculate the weight for each interval

The weight w_i for each interval is defined as the product of the inverse of the interval width and the representative value. This approach considers both the interval's precision and its magnitude:

$$w_i = \frac{r_i}{\text{width}_i + \epsilon} = \frac{r_i}{\bar{x}_i - \underline{x}_i + \epsilon} \quad (4)$$

where ϵ is a small positive constant (e.g., 10^{-5}) added to prevent division by zero when dealing with zero-width intervals.

Step 4: Normalize the weights

To ensure that the weights sum to 1, they are normalized as follows:

$$\hat{w}_i = \frac{w_i}{\sum_{j=1}^n w_j} \quad (5)$$

Step 5: Calculate the resulting interval

The final aggregated interval is computed according to Equation (1) as the weighted arithmetic mean of the $x_i = [\underline{x}_i, \bar{x}_i]$ using the normalized weights determined as explained above.

3.4. Feature Selection and Final Model Training

Features are ranked in descending order based on their calculated scores. Given a feature F_k , we assume m intervals of importance $y_k = [\underline{y}_k, \bar{y}_k]$ for $k = 1, 2, \dots, m$ are available.

The objective is to derive a score that links both the feature's importance (represented by the interval's center) and its stability (represented by the interval's width). This score will then be used to rank the features in terms of their overall significance.

Step 1: Interval Center and Width Calculation

For each interval $\mathbf{y}_k$,
calculate the center_k as the midpoint of the interval:

$$\text{center}_k = \frac{y_k + \bar{y}_k}{2}. \quad (6)$$

Calculate the width_k of interval $\mathbf{y}_k = [y_k, \bar{y}_k]$ as the difference between the upper and lower bounds (see Equation (3)), i.e., $\text{width}_k = \bar{y}_k - y_k$.

Step 2: Normalization of Center and Width

To ensure comparability across features, we normalize the center and width values to the range $[0, 1]$:

Normalize the center:

$$\text{norm_center}_k = \frac{\text{center}_k - \min(\text{center})}{\max(\text{center}) - \min(\text{center})} \quad (7)$$

where $\min(\text{center})$ and $\max(\text{center})$ are the minimum and maximum center values across all intervals and all features, respectively.

Normalize the width:

$$\text{norm_width}_k = 1 - \frac{\text{width}_k - \min(\text{width})}{\max(\text{width}) - \min(\text{width})} \quad (8)$$

where $\min(\text{width})$ and $\max(\text{width})$ are the minimum and maximum width values across all intervals and all features, respectively. Note that we use $1 - \text{normalized width}$ so that smaller widths result in larger values, reflecting higher stability.

Step 3: Score Calculation

The score_k for each interval is computed as a weighted combination of the normalized center and the normalized width. The weights α and β dictate the relative influence of these two components. We investigate several averaging methods to combine these components:

Weighted arithmetic mean

$$\text{score}_k^{(A)} = \alpha \cdot \text{norm_center}_k + \beta \cdot \text{norm_width}_k \quad (9)$$

Weighted geometric mean

$$\text{score}_k^{(G)} = (\text{norm_center}_k)^\alpha \cdot (\text{norm_width}_k)^\beta \quad (10)$$

Weighted harmonic mean

$$\text{score}_k^{(H)} = \frac{\alpha + \beta}{\frac{\alpha}{\text{norm_center}_k + \epsilon} + \frac{\beta}{\text{norm_width}_k + \epsilon}} \quad (11)$$

where ϵ is a small positive constant (e.g., 10^{-5}) to prevent division by zero.

These averaging methods offer different sensitivities to the magnitude of the normalized center and width:

- Arithmetic mean—provides a linear balance between importance and stability.
- Geometric mean—exhibits higher sensitivity to smaller values (a near-zero value in either component will significantly lower the score).

- Harmonic mean—amplifies the impact of small values even further than the geometric mean, making it suitable when low stability or importance will drastically reduce the score.

Step 4: Feature Ranking

Finally, the features are ranked in descending order according to their score_k values. Features with higher scores are considered more important, as they exhibit both high relevance (a large central value) and high stability (a small width, resulting in a large normalized width). A detailed numerical example illustrating the proposed algorithm can be found in [13].

Subsequently, a subset of features is selected according to a chosen selection strategy. This strategy may involve choosing features whose scores exceed the average or median score. The algorithm also ensures that a minimum number of features are retained. Finally, a new IVWFR model is trained on the entire dataset using only the selected subset of features. This final model can then be used to perform predictions on previously unseen data.

3.5. Hyperparameters and Their Optimization Using Optuna

The IVWFR algorithm uses the following hyperparameters:

- `n_splits`—number of folds for stratified k -fold cross-validation.
- `fold_size`—desired proportion of the dataset used for each training fold.
- `random_state`—seed for the random number generator, ensuring reproducibility.
- `selection_method`—method for selecting the features mean and median.
- `alpha`—weight given to interval's center in feature score (average importance).
- `representative`—representative point in interval: left, right or center.
- `mean_type`—type of mean to combine center and width: arithmetic, geometric or harmonic.

The performance of the IVWFR algorithm depends on its hyperparameter configuration. To automate the search for optimal settings, the Optuna framework [18] is employed. At the core of the optimization process lies the objective function, which receives a set of hyperparameters as input and returns a performance score that Optuna seeks to maximize.

For each set of hyperparameters (a trial), Optuna uses its suggest methods to sample values for each parameter within predefined ranges. It then performs 3-fold stratified cross-validation on the input dataset. In each fold, the IVWFR model is trained on the training portion using the current hyperparameter values, and its feature ranking performance is evaluated on the test portion using a custom metric.

The custom metric is designed to assess how well the model ranks features according to their known types. It assigns a higher score when truly relevant features are ranked at the top, redundant features appear near the top, and irrelevant features are placed lower in the ranking. Specifically, the metric awards 1 point if a relevant feature appears among the top-ranked ones, 0.5 points if it is ranked slightly lower but still within the top set of relevant and redundant features, and 0.5 points if a redundant feature is included among the top-ranked relevant and redundant features (cf. [13]). The final score is normalized to the range $[0, 1]$, and the objective function computes the mean score across all three folds. Optuna then maximizes this average score during the optimization process.

After the optimization is complete, Optuna returns the hyperparameter combination that achieves the highest average metric value. These optimal hyperparameters are subsequently used to train the final IVWFR model for feature selection on the entire dataset.

3.6. Performance Metrics

To evaluate experiments, the following classification metrics are used: accuracy, balanced accuracy, recall, precision, and F1 score.

Accuracy [19] is calculated as the number of accurate predictions divided by the total amount of predictions and represents the overall accuracy of the model.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Balanced accuracy provides a more informative evaluation of model performance than conventional accuracy, particularly when dealing with imbalanced class distributions. It is defined as the arithmetic mean of the recall scores for each class [20]. This metric allows the performance on each class to be assessed independently while accounting for class imbalance, making it especially suitable for datasets with unequal class representation.

Let

C —total number of classes,

i —each class's index,

TP_i —true positives for class i ,

FN_i —false negatives for class i ,

FP_i —false positives for class i .

The following formula is used for calculating multiclass balanced accuracy:

$$\text{Balanced Accuracy} = \frac{1}{C} \sum_{i=1}^C \frac{TP_i}{TP_i + FN_i}.$$

Recall [20] is defined as the ratio of true positive predictions to the total number of actual positive instances. It represents the proportion of positive cases that the model correctly identifies. Several variants of this metric exist for multiclass classification, including micro, macro, and weighted versions. In this study, we employ the weighted recall, which calculates the average recall across all classes by weighting each class according to its relative frequency in the dataset:

$$\text{Weighted Recall} = \sum_{i=1}^C \left(\frac{TP_i}{TP_i + FN_i} \cdot \frac{\text{Instances of class } i}{\text{Total instances}} \right).$$

Another metric used in the experiments is weighted precision. It is defined as the ratio of true positive predictions to the total number of true positive and false positive predictions. Precision [21] represents the proportion of positive predictions that are truly correct. The average precision across all classes is computed using weighted averaging, where the contribution of each class is weighted according to its relative frequency in the dataset:

$$\text{Weighted Precision} = \sum_{i=1}^C \left(\frac{TP_i}{TP_i + FP_i} \cdot \frac{\text{Instances of class } i}{\text{Total instances}} \right).$$

Finally, the F1 score [22] is defined as the harmonic mean of the model's precision and recall. Both precision and recall contribute equally to this metric. A model achieving a perfect F1 score demonstrates both perfect precision and perfect recall, meaning that it makes no false positive or false negative predictions. In this study, we use the weighted F1 score, which accounts for class imbalance by computing the class-wise F1 scores and

averaging them according to the number of samples in each class. The formula for the weighted F1 score is given below:

$$\text{Weighted F1 score} = \frac{2 \cdot \text{Weighted Precision} \cdot \text{Weighted Recall}}{\text{Weighted Precision} + \text{Weighted Recall}}$$

4. Datasets

In the experiments, two types of datasets are used: real datasets and synthetic datasets. Characteristics of these datasets are presented in Table 1 and Table 2, respectively.

The synthetic datasets used in this study were obtained from the SynthSelect toolkit [23]. This toolkit provides a collection of ten benchmark datasets, each containing predefined relevant, redundant, and irrelevant features. These datasets exhibit varying levels of complexity, making them particularly suitable for evaluating feature selection algorithms. Among them, the Double Spiral and 5D XOR datasets are considered the most challenging in terms of correctly identifying feature types (see Table 3 and Figures 11 and 12 in [23]). All datasets provided by the SynthSelect toolkit were used in our experiments, and to avoid potential overfitting, cross-validation was applied during both feature selection and performance evaluation phases.

In addition, five real-world datasets were employed: Climate, Estimation of Obesity Levels, Iranian Churn, and Sonar from the UCI Machine Learning Repository, as well as the Water Potability dataset from Kaggle. Using both synthetic and real datasets enables a comprehensive assessment of the proposed method's effectiveness.

All experiments were conducted in Python 3.13 using open-source libraries, including NumPy 2.3.3, pandas 2.3.2, SciPy 1.16.2, scikit-learn 1.7.2, XGBoost 3.0.5, Optuna 4.5.0, and matplotlib 3.10.6.

Table 1. Summary of real datasets.

Dataset	Rows	Features	Classes (Ratios)
Climate	540	18	2 (0.09, 0.91)
Estimation of Obesity	2111	16	7 (0.13, 0.14, 0.17, 0.14, 0.15, 0.14, 0.14)
Iranian Churn	3150	13	2 (0.84, 0.16)
Sonar	208	60	2 (0.53, 0.47)
Water Potability	3276	9	2 (0.61, 0.39)

Table 2. Summary of synthetic datasets. In this case, Rel, Red, and Irr denote the ground truth number of relevant, redundant, and irrelevant features in the dataset, respectively.

Dataset	Rows	Features	Classes (Ratios)	Rel	Red	Irr
10Class Multicut	500	126	10 (0.09, 0.06, 0.04, 0.04, 0.14, 0.18, 0.14, 0.14, 0.09, 0.09)	6	20	100
4D AND	100	113	2 (0.56, 0.44)	4	9	100
5Class Multicut	500	126	5 (0.03, 0.05, 0.05, 0.17, 0.70)	6	20	100
5D XOR	100	115	2 (0.49, 0.51)	5	10	100
Cone	400	123	2 (0.54, 0.46)	3	20	100
Double Spiral	200	303	2 (0.50, 0.50)	3	37	263
Hypersphere 3D	400	123	2 (0.52, 0.48)	3	20	100
Trig	200	57	2 (0.56, 0.43)	2	5	50
Yinyang	601	62	2 (0.51, 0.49)	2	10	50
y = x	200	122	2 (0.49, 0.51)	2	20	100

5. Results

This section presents a comparative analysis of the performance of various feature selection models on both real-world and synthetic datasets. We begin with a visual ex-

amination of the models' performance, followed by a statistical evaluation assessing the significance of the obtained results.

5.1. Performance Analysis

Figures A1–A10 present the results for the ten synthetic datasets, which were designed to evaluate model behavior under controlled conditions. Figure 1 consolidates these results to illustrate overall performance trends.

Similarly, Figures A11–A15 show the performance of each model on the five real-world datasets, while Figure 2 summarizes the results obtained on this practical, non-ideal data.

The models are ranked in descending order according to their balanced accuracy. For most of the real-world datasets, the proposed IVWFR method achieved the best performance. This observation is further confirmed in Figure 2, which aggregates the results from all real datasets.

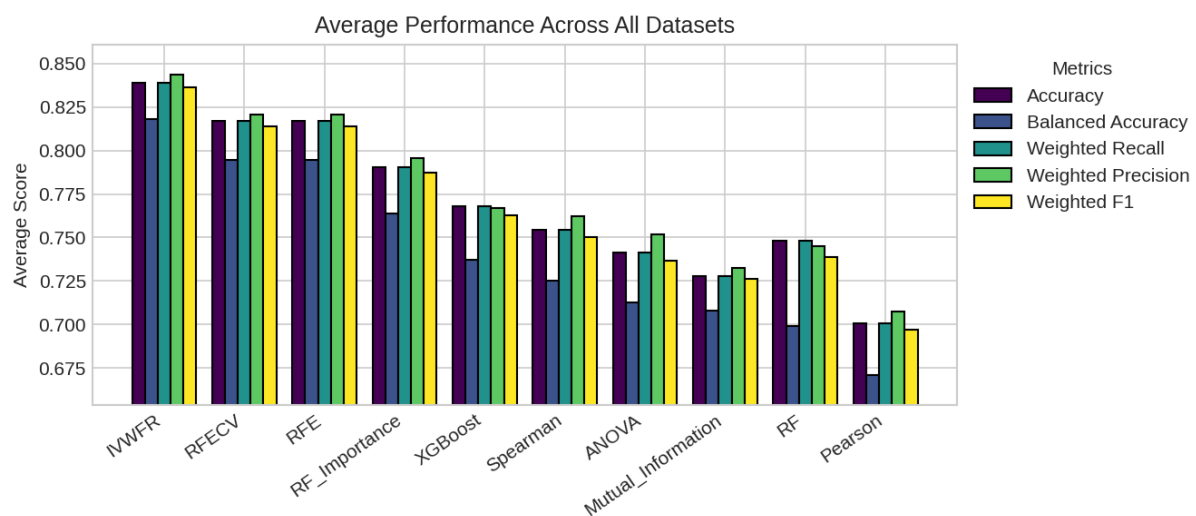


Figure 1. Summary of performance on synthetic datasets. This figure provides a consolidated view of the results across all synthetic datasets.



Figure 2. Summary of performance on real datasets. This figure provides a consolidated view of the results across all real datasets.

A similar trend is observed in the case of synthetic datasets. The proposed IVWFR method appears particularly effective for the Double Spiral and 5D XOR datasets, which—as previously mentioned—are among the most challenging due to the difficulty of correctly

identifying feature types. For these datasets, the performance differences in favor of IVWFR are the most pronounced. As shown in Figure 1, which consolidates all synthetic results, IVWFR consistently outperforms the competing methods.

5.2. Statistical Significance and Discussion

To assess the statistical significance of the performance differences, the Wilcoxon signed-rank test was conducted separately for the real and synthetic datasets. This non-parametric test compares pairs of models based on their balanced accuracy scores. The resulting p -values, corrected for multiple comparisons using the Bonferroni method, are reported in Tables 3 and 4.

For the real-world datasets (Table 4), performance varied across models; however, IVWFR demonstrated a statistically significant advantage over traditional methods such as ANOVA, Pearson, and Spearman correlation. Nonetheless, many models—including RFECV, RFE, and the baseline classifiers—did not differ significantly from one another, suggesting that for these datasets, the use of more sophisticated feature selection methods may not be crucial. The lack of significant differences between some methods on real data may result from reaching a “ceiling” in model quality, while IVWFR stands out with fewer selected features while maintaining the same effectiveness.

In contrast, the analysis of the synthetic datasets (Table 3) revealed more pronounced distinctions among the evaluated methods. In this case, the IVWFR model significantly outperformed nearly all competing approaches, indicating that the algorithm is particularly effective at identifying relevant features in datasets with complex and well-defined structures. As expected, the performance of RFE and RFECV was statistically indistinguishable, since RFE was configured to select the same number of features that RFECV determined as optimal.

Table 3. p -values for synthetic datasets from the Wilcoxon signed-rank test *.

	IVWFR	RFECV	ANOVA	Pearson	Spearman	RF-Imp.	Mutual Inf.	RFE	RF	XGBoost
IVWFR	-	0.041	2.2×10^{-7}	1.4×10^{-7}	1.3×10^{-6}	2.3×10^{-4}	2.1×10^{-6}	0.041	5.9×10^{-8}	2.1×10^{-6}
RFECV	0.041	-	5.3×10^{-5}	1.4×10^{-5}	0.008	0.038	6.1×10^{-4}	1	2.0×10^{-6}	1.6×10^{-4}
ANOVA	2.2×10^{-7}	5.3×10^{-5}	-	0.217	1	0.423	1	5.3×10^{-5}	1	1
Pearson	1.4×10^{-7}	1.4×10^{-5}	0.217	-	0.022	8.3×10^{-5}	1	1.4×10^{-5}	1	0.259
Spearman	1.3×10^{-6}	0.008	1	0.022	-	1	1	0.008	1	1
RF-Imp.	2.3×10^{-4}	0.038	0.423	8.3×10^{-5}	1	-	0.024	0.038	0.005	0.198
Mutual Inf.	2.1×10^{-6}	6.1×10^{-4}	1	1	1	0.024	-	6.1×10^{-4}	1	1
RFE	0.041	1	5.3×10^{-5}	1.4×10^{-5}	0.008	0.038	6.1×10^{-4}	-	2.0×10^{-6}	1.6×10^{-4}
RF	5.9×10^{-8}	2.0×10^{-6}	1	1	1	0.005	1	2.0×10^{-6}	-	0.331
XGBoost	2.1×10^{-6}	1.6×10^{-4}	1	0.259	1	0.198	1	1.6×10^{-4}	0.331	-

* RF-Imp.: RF Importance; Mutual Inf.: Mutual Information. The table shows Bonferroni-corrected p -values. Values in bold indicate statistical significance ($p < 0.05$).

Table 4. p -values for real datasets from the Wilcoxon signed-rank test *.

	IVWFR	RFECV	ANOVA	Pearson	Spearman	RF-Imp.	Mutual Inf.	RFE	RF	XGBoost
IVWFR	-	1	0.010	0.010	0.005	1	0.021	1	0.552	1
RFECV	1	-	0.114	0.114	0.292	1	0.321	1	1	1
ANOVA	0.010	0.114	-	1	1	0.008	1	0.114	1	0.064
Pearson	0.010	0.114	1	-	1	0.008	0.834	0.114	1	0.064
Spearman	0.005	0.292	1	1	-	0.029	0.999	0.292	1	0.040
RF-Imp.	1	1	0.008	0.008	0.029	-	0.052	1	1	1
Mutual Inf.	0.021	0.321	1	0.834	0.999	0.052	-	0.321	1	0.097
RFE	1	1	0.114	0.114	0.292	1	0.321	-	1	1
RF	0.552	1	1	1	1	1	1	1	-	1
XGBoost	1	1	0.064	0.064	0.040	1	0.097	1	1	-

* RF-Imp.: RF Importance; Mutual Inf.: Mutual Information. The table shows Bonferroni-corrected p -values. Values in bold indicate statistical significance ($p < 0.05$).

5.3. Number of Features Selected

Before analyzing feature overlap, we first examine the number of features selected by each method, as this provides essential context for interpreting the Jaccard similarity scores. Tables 5 and 6 summarize the feature counts for the real-world and synthetic datasets, respectively.

Several key patterns emerge from these results. By experimental design, most conventional methods (from RFECV to RFE) select the same number of features, as this count is fixed to the optimum determined by RFECV. In contrast, the IVWFR method frequently selects a different number of features, reflecting its distinct selection strategy. Finally, the RF and XGBoost models show the highest feature counts because they serve as baseline classifiers that utilize all available features. This variation in the number of selected features is an important factor influencing the subsequent overlap analysis. The number of features for comparative methods is set at the optimal level determined by RFECV to enable a reliable comparison with an identical feature space size. In order to assess sensitivity, we conduct an additional analysis in which the number of selected features is changed and the quality of the models is measured. The results of this analysis for individual datasets are presented in the graphs in Figures A16–A30 in Appendix A.

Table 5. Number of features selected for each synthetic dataset and each method. RF-Imp. is an abbreviation for RF Importance and Mutual Inf. for Mutual Information.

Dataset	IVWFR	RFECV	ANOVA	Pearson	Spearman	RF Imp.	Mutual Inf.	RFE	RF	XGBoost
5Class Multicut	8	7	7	7	7	7	7	7	126	126
10Class Multicut	39	11	11	11	11	11	11	11	126	126
Trig	5	5	5	5	5	5	5	5	57	57
Hypersphere 3D	19	95	95	95	95	95	95	95	123	123
4D AND	25	4	4	4	4	4	4	4	113	113
Double Spiral	152	227	227	227	227	227	227	227	303	303
Cone	26	25	25	25	25	25	25	25	123	123
5D XOR	40	35	35	35	35	35	35	35	115	115
y = x	23	14	14	14	14	14	14	14	122	122
Yinyang	11	7	7	7	7	7	7	7	62	62

Table 6. Number of features selected for each real dataset and each method. RF-Imp. is an abbreviation for RF Importance and Mutual Inf. for Mutual Information.

Dataset	IVWFR	RFECV	ANOVA	Pearson	Spearman	RF Imp.	Mutual Inf.	RFE	RF	XGBoost
estimation_of_obesity	5	2	2	2	2	2	2	2	16	16
water_potability	5	4	4	4	4	4	4	4	9	9
iranian_churn	13	11	11	11	11	11	11	11	13	13
sonar	30	47	47	47	47	47	47	47	60	60
climate	5	4	4	4	4	4	4	4	18	18

Simultaneously with the analysis of classification quality, in the case of synthetic data, an analysis of ground truth feature identification was performed. The TPRel, TPRed and TPIrr evaluation measures were used for this purpose, which stand for true positive relevant, true positive redundant, and true positive irrelevant [13]. In addition, the TCC measure, or total correct classification, also defined in [13], was used, combining TPRel and TPRed as the quality of identification of all relevant features in the set. The results of individual measures are presented in Figure 3, and a summary of these results for the six best feature selection methods is presented in Figure 4.

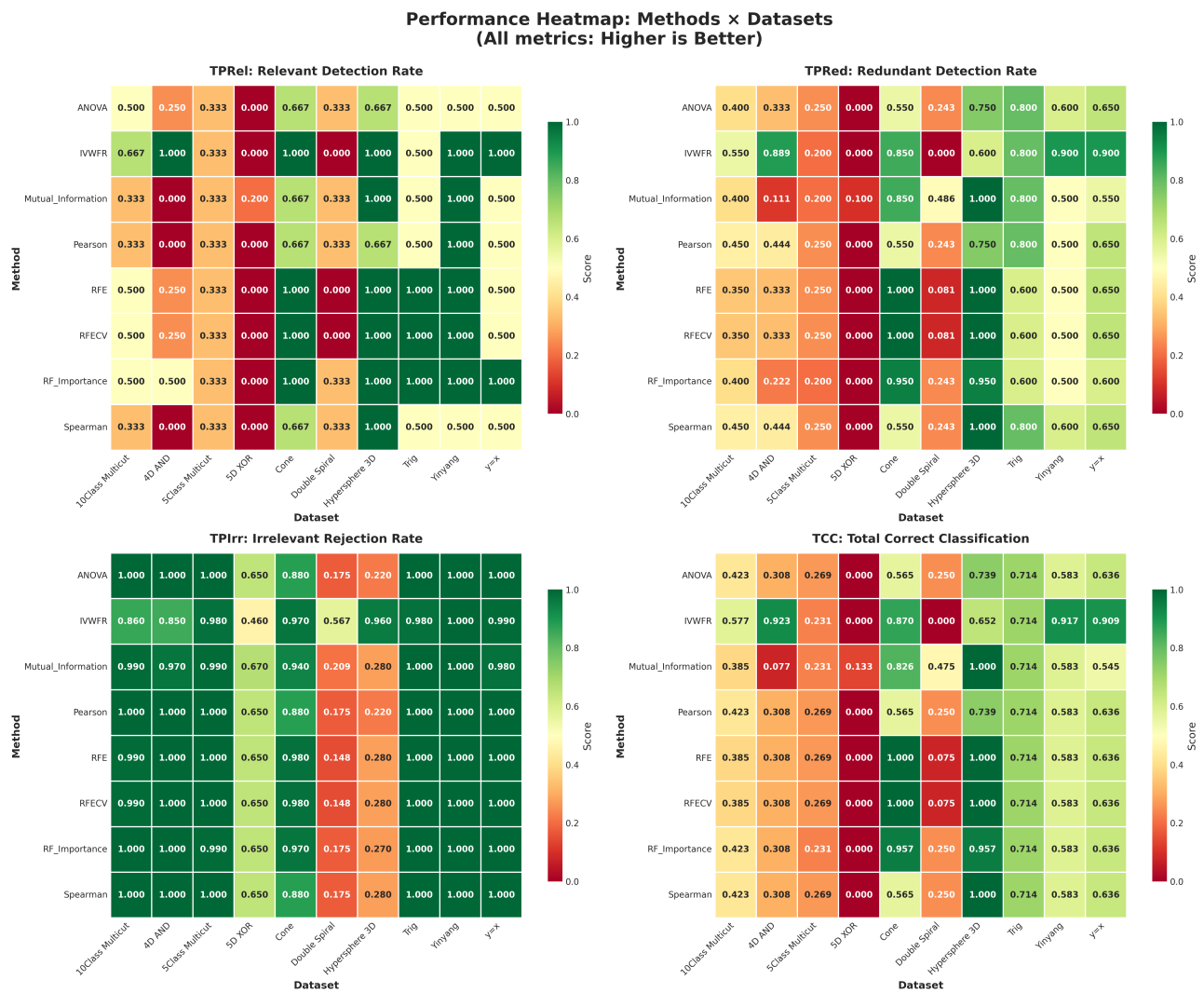


Figure 3. Results of ground truth feature identification for synthetic datasets in the context of the feature selection method used. Individual graphs show the identification of relevant (TPRel), redundant (TPRed), and irrelevant (TPIrr) features, and total correct classification (TCC), i.e., the correct identification of relevant and redundant features [13].

5.4. Feature Selection Overlap Analysis

Beyond performance metrics, we also examined the degree of agreement among feature selection models regarding which features they identified as important. The Jaccard similarity index was used to quantify the overlap between the feature subsets selected by each pair of models, where a value of 1.0 indicates identical sets and a value of 0.0 indicates no shared features.

5.4.1. Overlap on Synthetic Datasets

The analysis of the synthetic datasets (Tables 7–16) makes the algorithmic differences between the models more apparent. Notably, IVWFR often selects a different number of features compared to the other methods, which are constrained by the RFECV output. This confirms that IVWFR relies on a fundamentally different principle for identifying the optimal feature set. While certain groups of methods (e.g., statistical filters) still show some degree of agreement, the overall similarity is lower than that observed for the real-world datasets. Consequently, the complex structures of the synthetic data are more effective in revealing the distinctive behavior of each feature selection strategy.

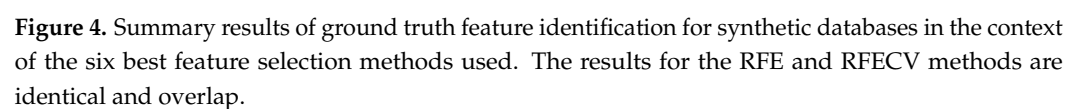


Table 7. Jaccard similarity on the 5Class Multicut dataset.

[illegible]

Table 8. Jaccard similarity on the 10Class Multicut dataset.[illegible]**Table 9.** Jaccard similarity on the Trig dataset.[illegible]**Table 10.** Jaccard similarity on the Hypersphere 3D dataset.[illegible]**Table 11.** Jaccard similarity on the 4D AND dataset.[illegible]

Table 16. Jaccard similarity on the Yinyang dataset.

[illegible]

5.4.2. Overlap on Real Datasets

For the real-world datasets, the Jaccard similarity matrices (Tables 17–21) show that methods based on Random Forest feature importance (RFECV, RF-Imp., and RFE) often exhibit perfect or near-perfect similarity, as they rely on the same underlying feature ranking. Similarly, filter-based approaches such as ANOVA and Pearson correlation frequently produce comparable feature subsets. In contrast, the IVWFR model consistently demonstrates a more distinct selection pattern, underscoring its unique selection strategy. The lower agreement between the IVWFR and RFE/RFECV methods in feature selection results from a different selection philosophy—IVWFR focuses on stability and interval significance, thus selecting a smaller but more relevant set of features, which translates into classification efficiency.

Table 17. Jaccard similarity on the Climate dataset.

[illegible]**Table 18.** Jaccard similarity on the Iranian Churn dataset.[illegible]

Table 19. Jaccard similarity on the Sonar dataset.

	IVWFR	RFECV	ANOVA	Pearson	Spearman	RF-Imp.	Mutual Inf.	RFE	RF	XGBoost
IVWFR	1.000	0.638	0.510	0.510	0.540	0.638	0.426	0.638	0.500	0.500
RFECV	0.638	1.000	0.741	0.741	0.741	0.774	0.621	1.000	0.783	0.783
ANOVA	0.510	0.741	1.000	1.000	0.918	0.709	0.679	0.741	0.783	0.783
Pearson	0.510	0.741	1.000	1.000	0.918	0.709	0.679	0.741	0.783	0.783
Spearman	0.540	0.741	0.918	0.918	1.000	0.741	0.649	0.741	0.783	0.783
RF-Imp.	0.638	0.774	0.709	0.709	0.741	1.000	0.649	0.774	0.783	0.783
Mutual Inf.	0.426	0.621	0.679	0.679	0.649	0.649	1.000	0.621	0.783	0.783
RFE	0.638	1.000	0.741	0.741	0.741	0.774	0.621	1.000	0.783	0.783
RF	0.500	0.783	0.783	0.783	0.783	0.783	0.783	0.783	1.000	1.000
XGBoost	0.500	0.783	0.783	0.783	0.783	0.783	0.783	0.783	1.000	1.000

Table 20. Jaccard similarity on the Water Potability dataset.

	IVWFR	RFECV	ANOVA	Pearson	Spearman	RF-Imp.	Mutual Inf.	RFE	RF	XGBoost
IVWFR	1.000	0.800	0.500	0.500	0.500	0.800	0.500	0.800	0.556	0.556
RFECV	0.800	1.000	0.333	0.333	0.333	1.000	0.600	1.000	0.444	0.444
ANOVA	0.500	0.333	1.000	1.000	1.000	0.333	0.143	0.333	0.444	0.444
Pearson	0.500	0.333	1.000	1.000	1.000	0.333	0.143	0.333	0.444	0.444
Spearman	0.500	0.333	1.000	1.000	1.000	0.333	0.143	0.333	0.444	0.444
RF-Imp.	0.800	1.000	0.333	0.333	0.333	1.000	0.600	1.000	0.444	0.444
Mutual Inf.	0.500	0.600	0.143	0.143	0.143	0.600	1.000	0.600	0.444	0.444
RFE	0.800	1.000	0.333	0.333	0.333	1.000	0.600	1.000	0.444	0.444
RF	0.556	0.444	0.444	0.444	0.444	0.444	0.444	0.444	1.000	1.000
XGBoost	0.556	0.444	0.444	0.444	0.444	0.444	0.444	0.444	1.000	1.000

Table 21. Jaccard similarity on the Estimation of Obesity dataset.

	IVWFR	RFECV	ANOVA	Pearson	Spearman	RF-Imp.	Mutual Inf.	RFE	RF	XGBoost
IVWFR	1.000	0.400	0.167	0.167	0.167	0.400	0.400	0.400	0.312	0.312
RFECV	0.400	1.000	0.333	0.333	0.333	1.000	0.333	1.000	0.125	0.125
ANOVA	0.167	0.333	1.000	0.333	0.333	0.333	0.333	0.333	0.125	0.125
Pearson	0.167	0.333	0.333	1.000	1.000	0.333	0.333	0.333	0.125	0.125
Spearman	0.167	0.333	0.333	1.000	1.000	0.333	0.333	0.333	0.125	0.125
RF-Imp.	0.400	1.000	0.333	0.333	0.333	1.000	0.333	1.000	0.125	0.125
Mutual Inf.	0.400	0.333	0.333	0.333	0.333	0.333	1.000	0.333	0.125	0.125
RFE	0.400	1.000	0.333	0.333	0.333	1.000	0.333	1.000	0.125	0.125
RF	0.312	0.125	0.125	0.125	0.125	0.125	0.125	0.125	1.000	1.000
XGBoost	0.312	0.125	0.125	0.125	0.125	0.125	0.125	0.125	1.000	1.000

5.4.3. Feature Importance Stability Analysis

To evaluate the robustness and repeatability of the feature selection process, a stability analysis was performed across multiple data splits during 5-fold cross-validation. For each feature selection method, two complementary stability indices were computed to quantify the consistency of the selected feature subsets across different folds (see Figures A31 and A32 for synthetic datasets and Figures A33 and A34 for real-world datasets).

1. Jaccard Similarity Index (J):

$$J(A_i, A_j) = \frac{|A_i \cap A_j|}{|A_i \cup A_j|} \quad (12)$$

where A_i and A_j denote the sets of selected features obtained in two independent folds. The Jaccard index measures the proportion of common features between two selections and ranges from 0 (no overlap) to 1 (identical feature subsets).

2. Kuncheva Index (κ):

$$\kappa(A_i, A_j) = \frac{|A_i \cap A_j| \cdot N - d^2}{d \cdot (N - d)} \quad (13)$$

where N is the total number of available features and d is the number of features selected. The Kuncheva index corrects for random overlap between subsets, providing a more conservative estimate of stability.

Both indices were calculated for all pairs of folds and then averaged to obtain the overall stability of each feature selection method for synthetic (Figure 5) and real-world (Figure 6) datasets. Higher values of J and κ indicate greater robustness of the feature selection process to data perturbations and random partitioning.

Among the real-world datasets used, one set, Water Potability, contains approximately 4.86% missing data, which affects three of the nine features. Analysis of this issue (Figure A37) indicates that a 5% missing value rate for this database leads to a loss in classification quality from approximately 61.9% to only 61.7%, which confirms its applicability. We use the following strategy for imputing missing values: each missing feature value is replaced with the mean value of that feature. This strategy is applied before feature scaling and selection and is consistent across all feature selection methods. The summary results of the simulation analysis of missing values in the range of 5–20% for all sets are presented in Figure A37, while detailed values are presented in Tables A1–A5.

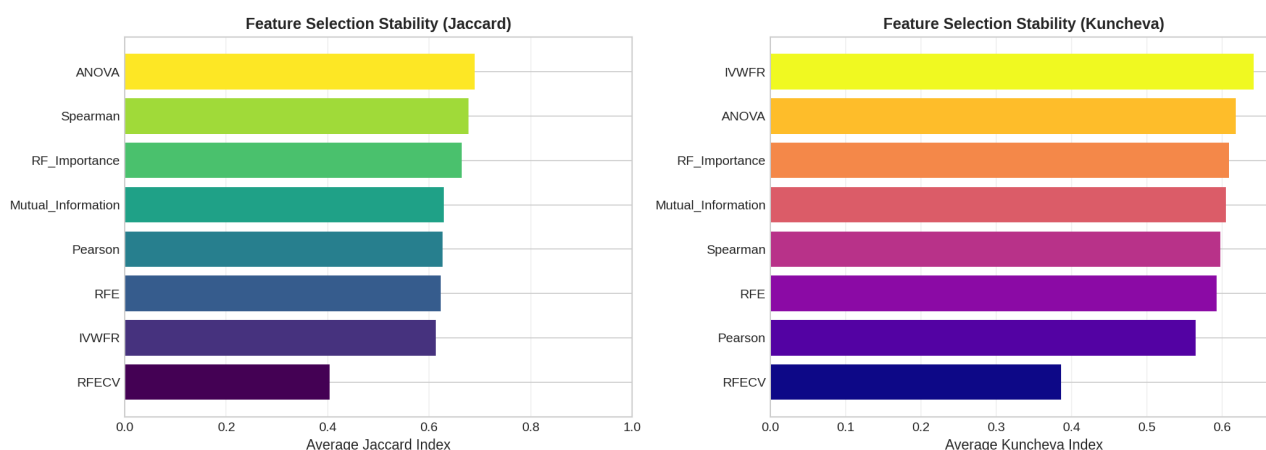


Figure 5. Average stability across synthetic datasets.

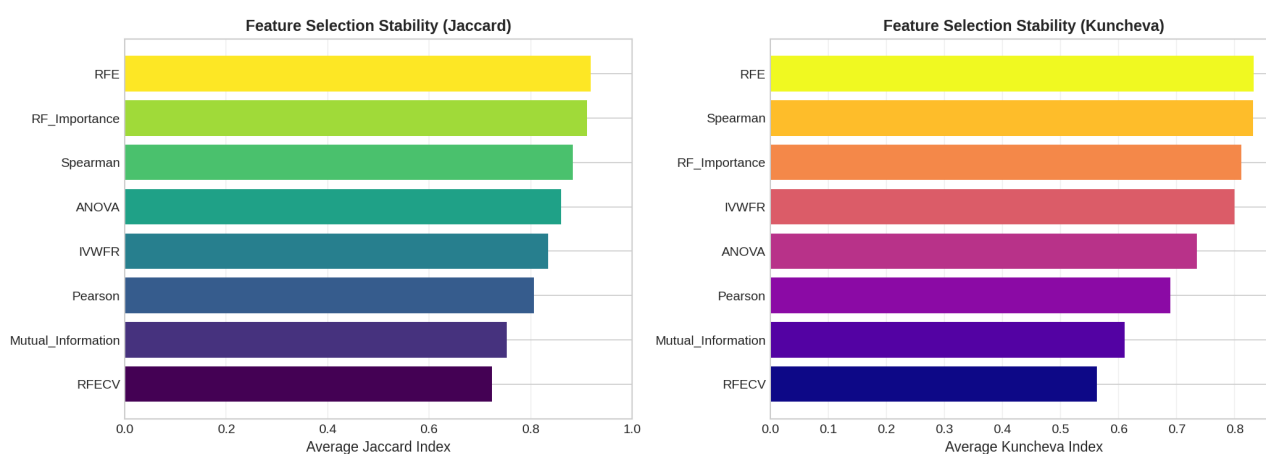


Figure 6. Average stability across real-world datasets.

6. Discussion

The research conclusions drawn from the results obtained are very optimistic. Analysis of synthetic data (Figures A1–A10) shows that the IVWFR algorithm achieves the best classification results among the selected methods, or ranks among the top results (e.g., Cone or Trig dataset). The summary classification chart (Figure 1) proves its effectiveness, where the results obtained by the IVWFR algorithm are significantly better than those of the other methods. Significant differences in the results achieved are noticeable, for example, for the 10Class Multicut dataset, for which the algorithm achieved the best classification results and, at the same time, identified 39 significant features compared to 11 features identified by the RFECV and RFE algorithms (see Table 5), which may indicate its sensitivity to less significant but important features that improve classification. A similarly sensitive result was obtained for the synthetic Double Spiral dataset, where the classification results were much better than those obtained using other methods (Figure A6). At the same time, the IVWFR algorithm identified only 152 significant features, while other methods identified between 227 and 303 features (Table 5). This may indicate the sensitivity of the algorithm in terms of removing irrelevant information noise from the dataset.

The results obtained for five real-world datasets are also very promising. The IVWFR algorithm achieved the best results or was among the best methods (Figures A11–A15). The summary results (Figure 2) also show that the IVWFR algorithm achieved the best classification results. At the same time, for example, for the Sonar dataset, it identified 30 significant features, while the other methods identified between 47 and 60 (Table 6) out of all 60 features in the dataset (Table 1), and the classification results were clearly better (Figure A14). Moreover, the *p*-values obtained for the IVWFR algorithm results also confirm its effectiveness for both synthetic data (Table 3) and real data (Table 4). The IVWFR method achieved comparable accuracy with significantly fewer features, confirming its effectiveness and ability to reduce complexity.

The results in Figures 5 and 6 clearly demonstrate that for the synthetic dataset the proposed IVWFR method achieves high feature selection stability among all compared approaches. The average Jaccard index reaches 0.61 and Kuncheva index reaches 0.64 (the best result) for synthetic datasets and 0.83 (Jaccard) and 0.80 (Kuncheva) for real-world datasets, indicating that IVWFR consistently selects a subset of most of the same features across different cross-validation folds.

However, simple statistical filters (ANOVA, Pearson, Spearman correlation) exhibit similar repeatability, suggesting the same sensitivity to data perturbations. In contrast, wrapper-based methods achieve moderate (RFE) or the worst (RFECV) stability but at the cost of considerably higher computational complexity (Figures A35 and A36).

The appropriate stability of IVWFR, both for synthetic and real-world data, can be attributed to its interval-valued aggregation mechanism, which smooths out the fluctuations in feature importance values obtained from multiple data partitions. By incorporating uncertainty intervals into the ranking process, the method effectively suppresses random noise and yields a more reproducible and interpretable feature set.

Future research directions will focus on developing the proposed approach, i.e., the use of innovative, adaptive methods for aggregating significance intervals; developing new weighting methods, nonlinear methods that take into account the influence of individual intervals; developing new methods for scoring significance, incorporating interval stability; and using optimization to determine feature contributions.

Author Contributions: Conceptualization, W.P. and U.B.; methodology, A.W. and U.B.; software, A.W.; validation, A.W. and W.P.; formal analysis, U.B.; investigation, A.W.; resources, A.W. and W.P.; data curation, A.W. and W.P.; writing—original draft preparation, A.W., W.P. and U.B.; writing—review and editing, U.B. and W.P.; visualization, A.W.; supervision, W.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository.

Acknowledgments: This work was partially supported by the Centre for Innovation and Transfer of Natural Sciences and Engineering Knowledge of University of Rzeszów, Poland.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

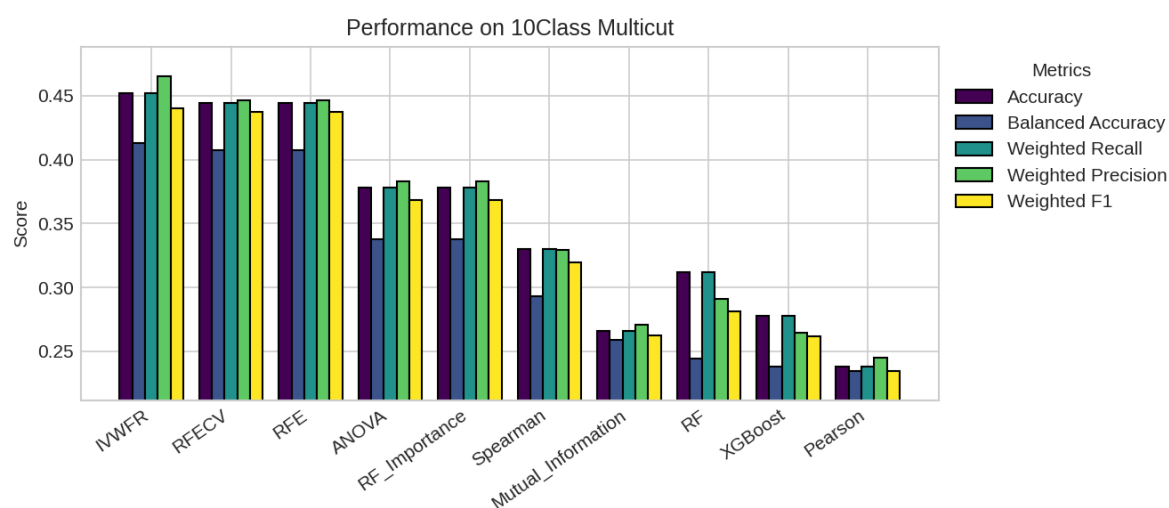


Figure A1. Performance on 10Class Multicut dataset.

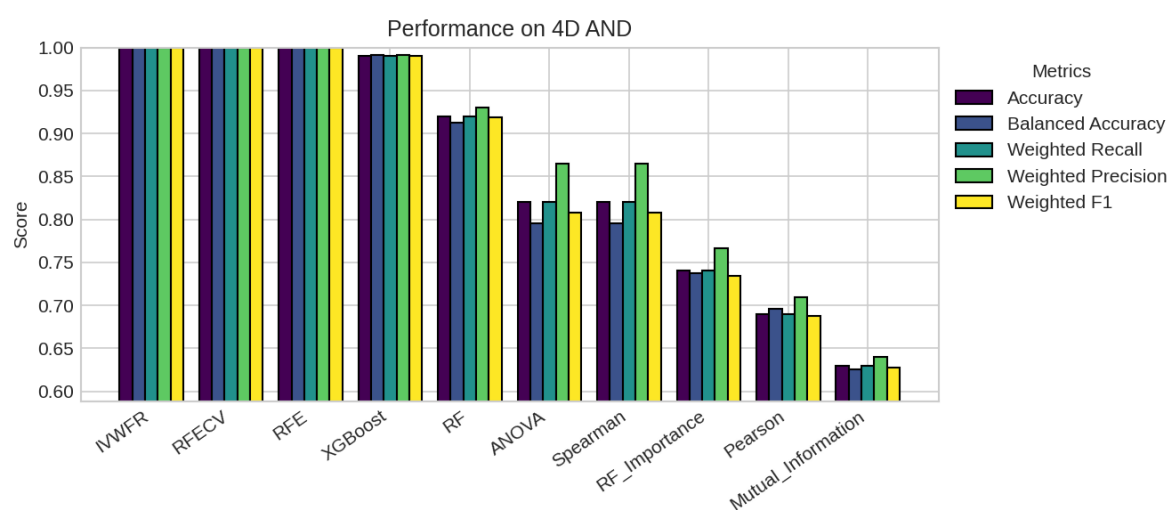


Figure A2. Performance on 4D AND dataset.



Figure A3. Performance on 5Class Multicut dataset.

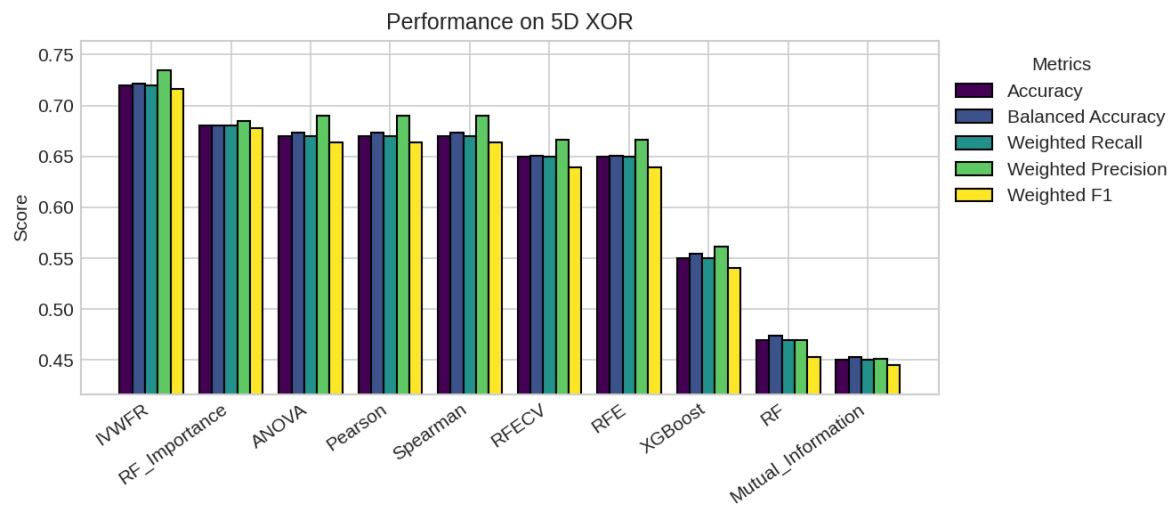


Figure A4. Performance on 5D XOR dataset.

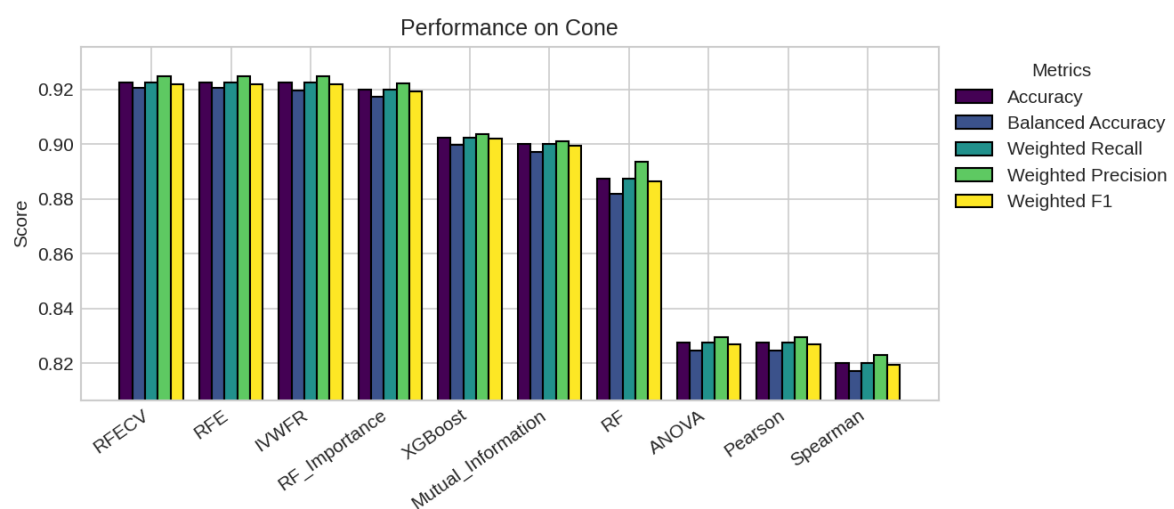


Figure A5. Performance on Cone dataset.

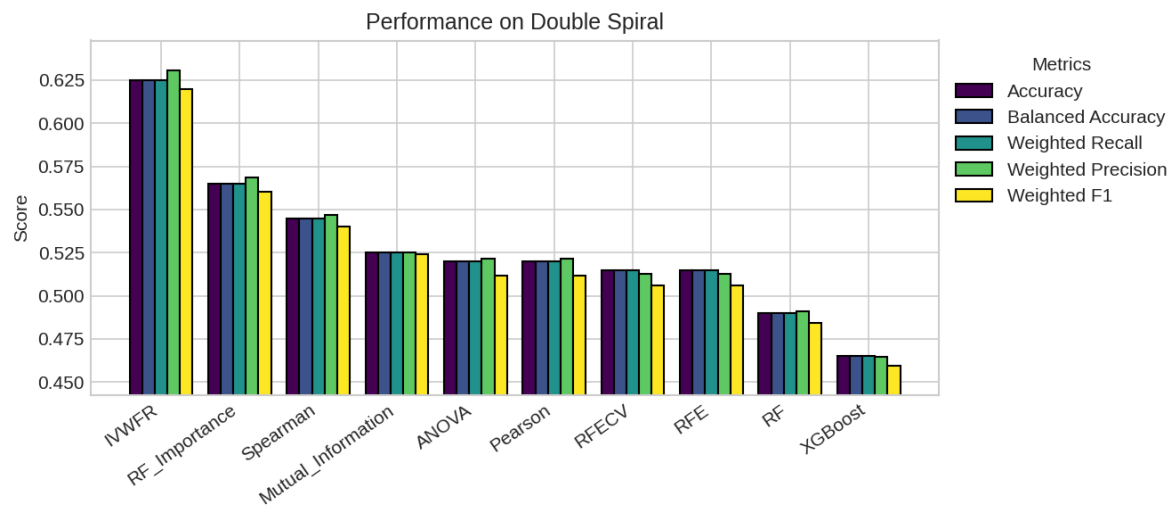


Figure A6. Performance on Double Spiral dataset.

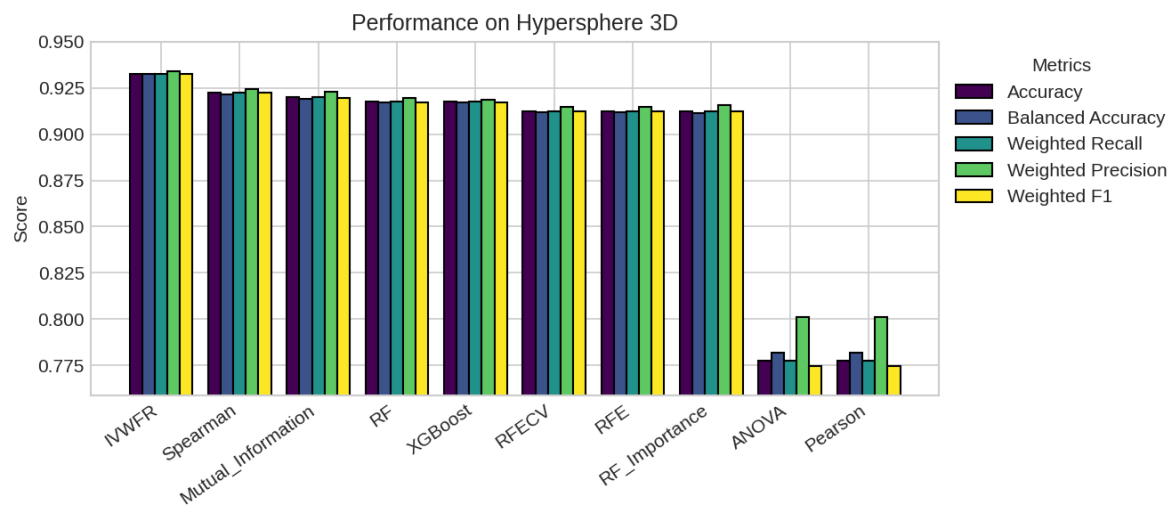


Figure A7. Performance on Hypersphere 3D dataset.

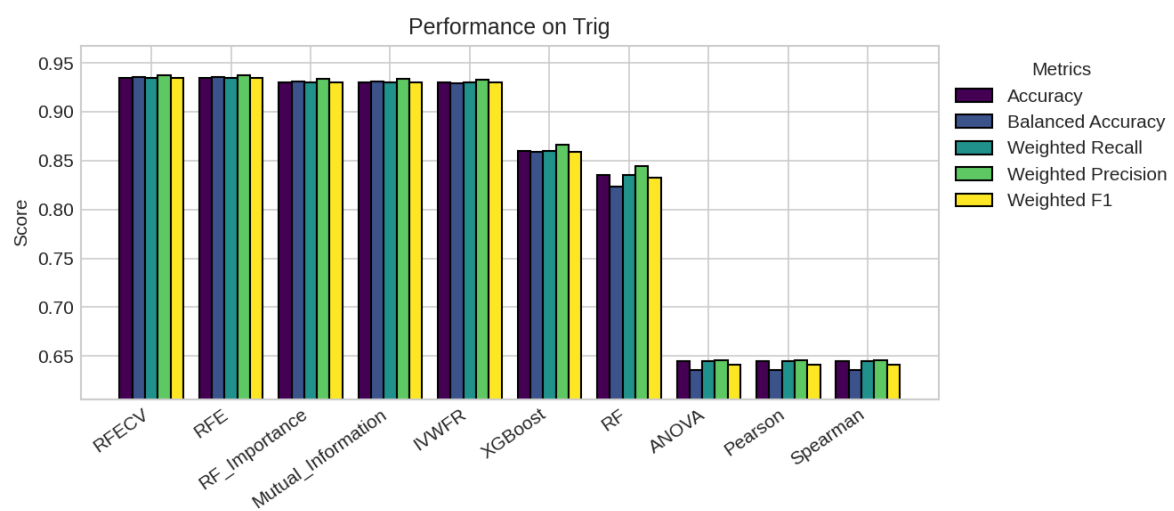


Figure A8. Performance on Trig dataset.

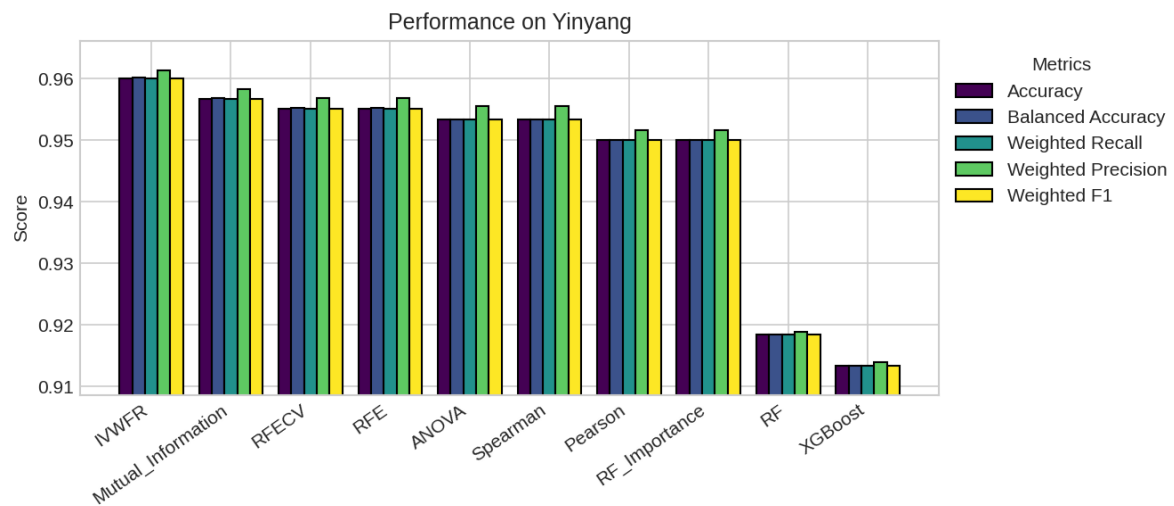


Figure A9. Performance on Yinyang dataset.

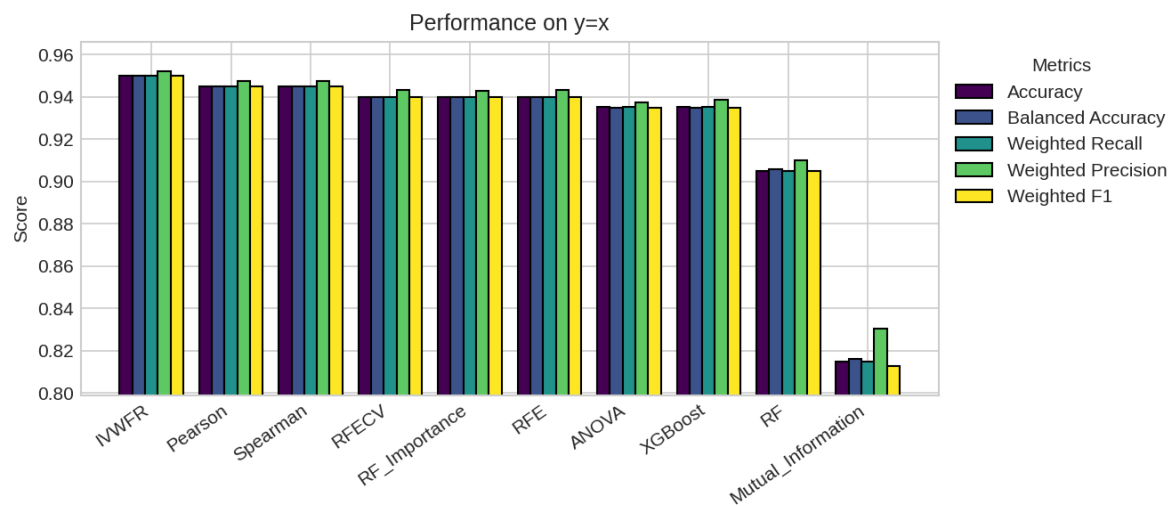


Figure A10. Performance on $y = x$ dataset.

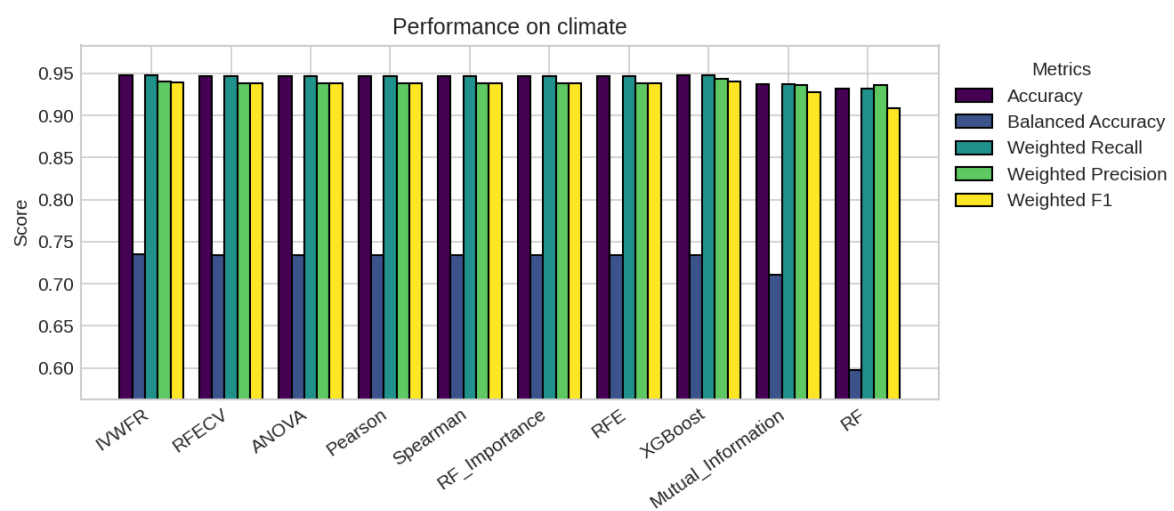


Figure A11. Performance on Climate dataset.

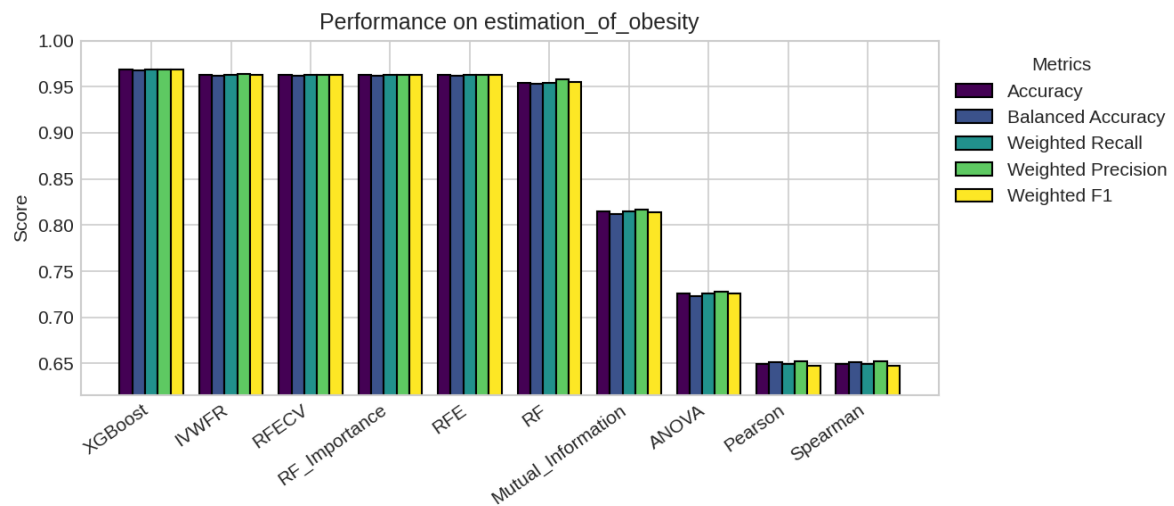


Figure A12. Performance on Estimation of Obesity Levels dataset.

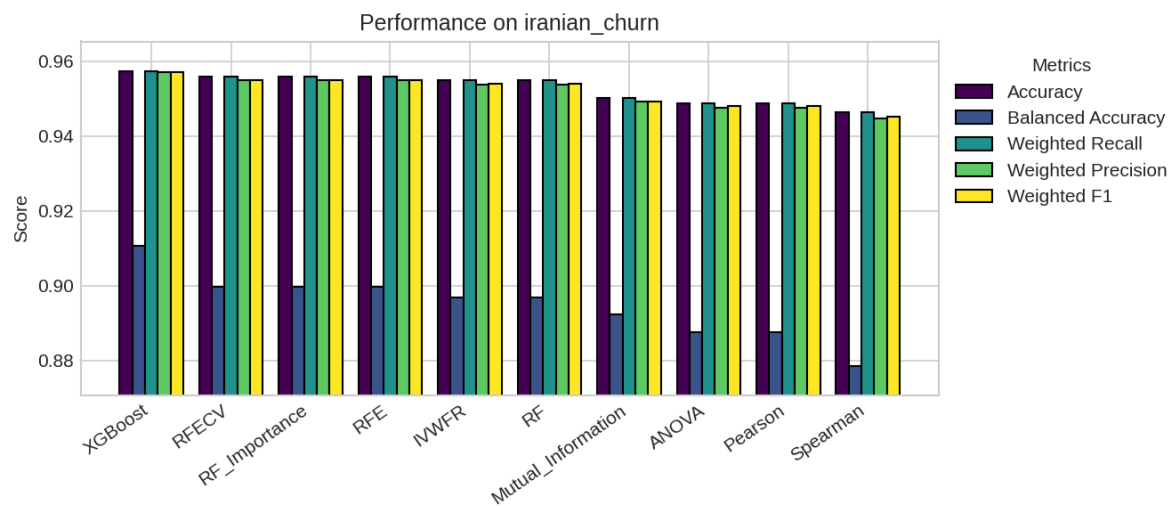


Figure A13. Performance on Iranian Churn dataset.

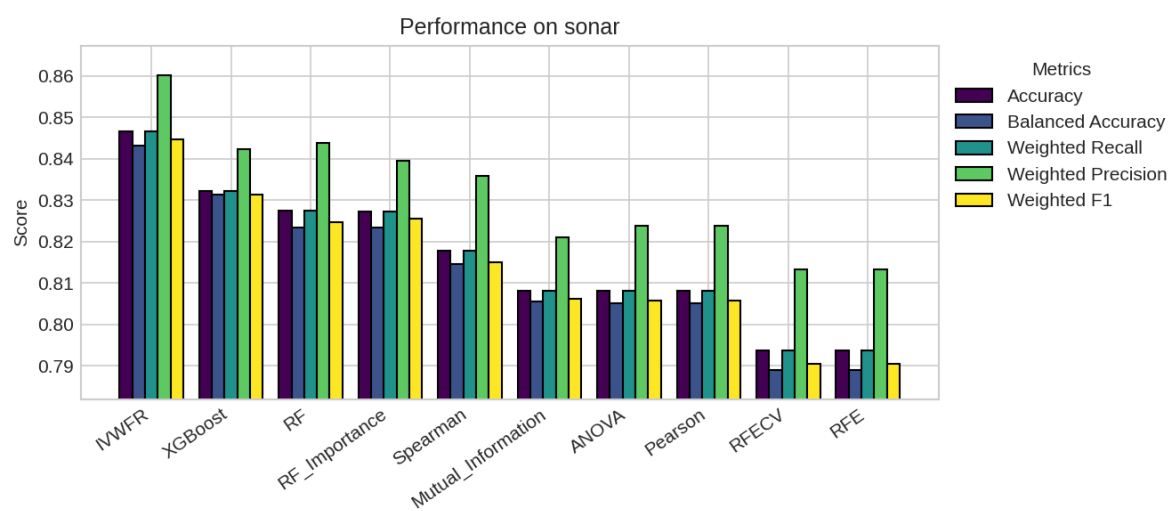


Figure A14. Performance on Sonar dataset.

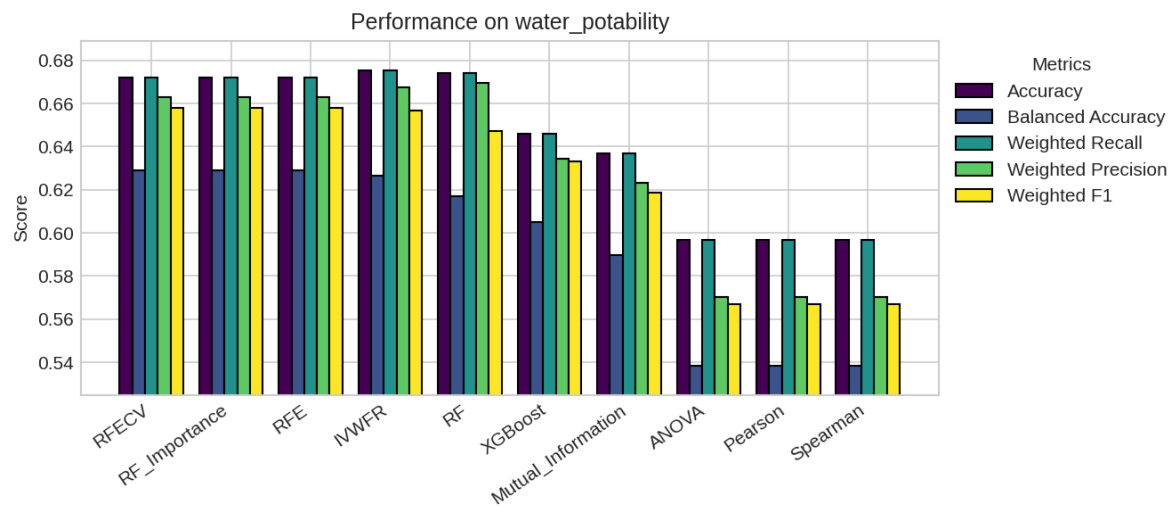


Figure A15. Performance on Water Potability dataset.

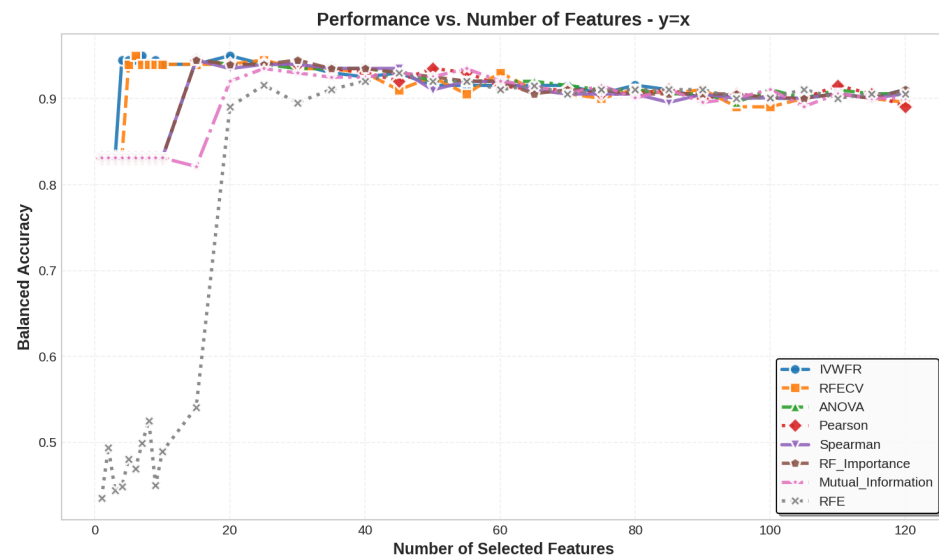


Figure A16. Performance compared to number of features using $y = x$ dataset.

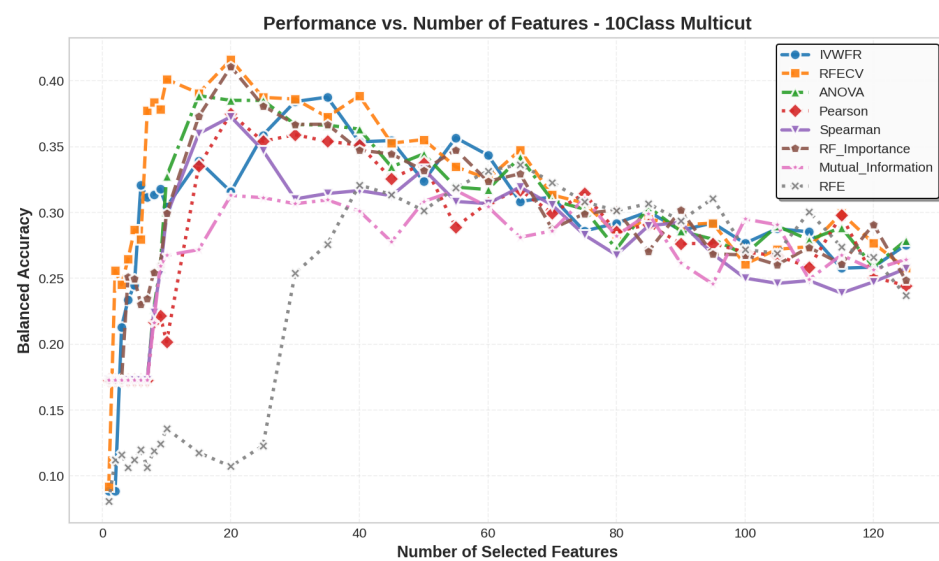


Figure A17. Performance compared to number of features using 10Class Multicut dataset.

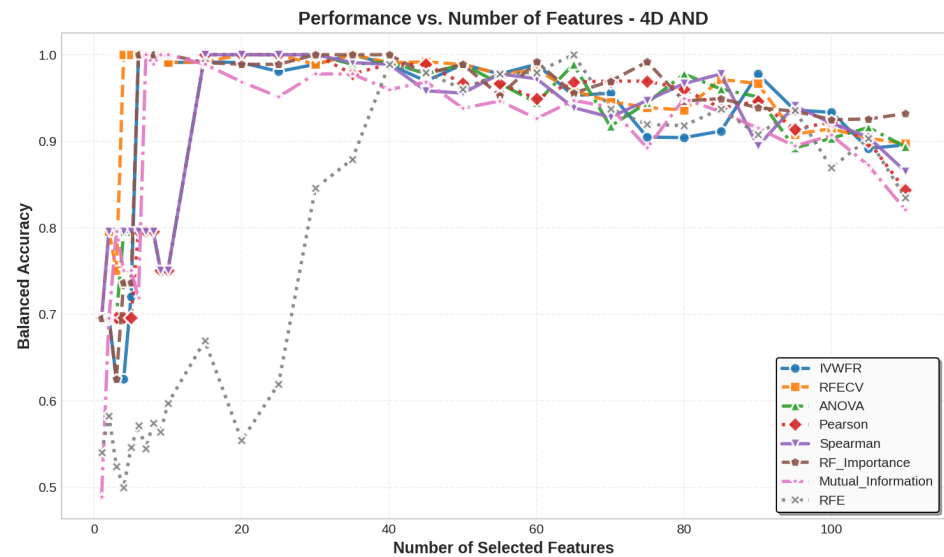


Figure A18. Performance compared to number of features using features_4D AND dataset.

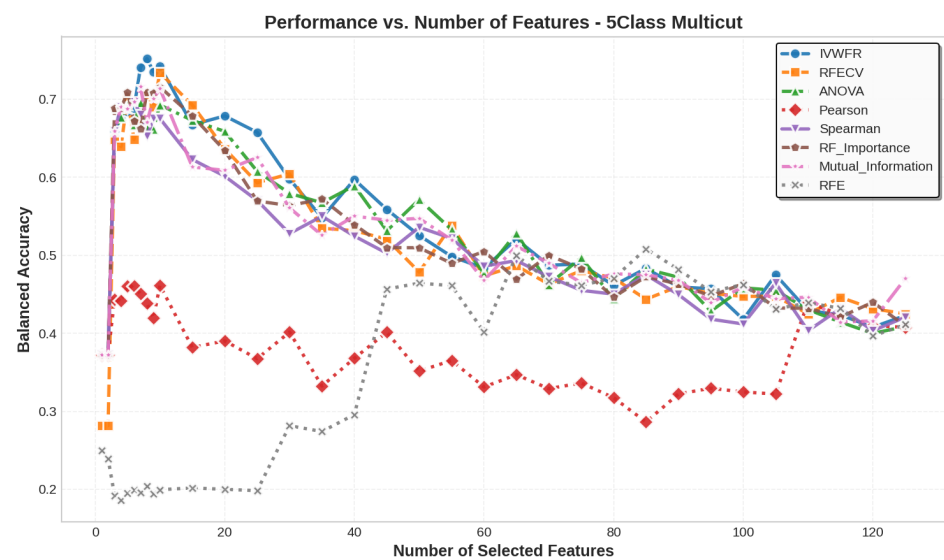


Figure A19. Performance compared to number of features using 5Class Multicut dataset.

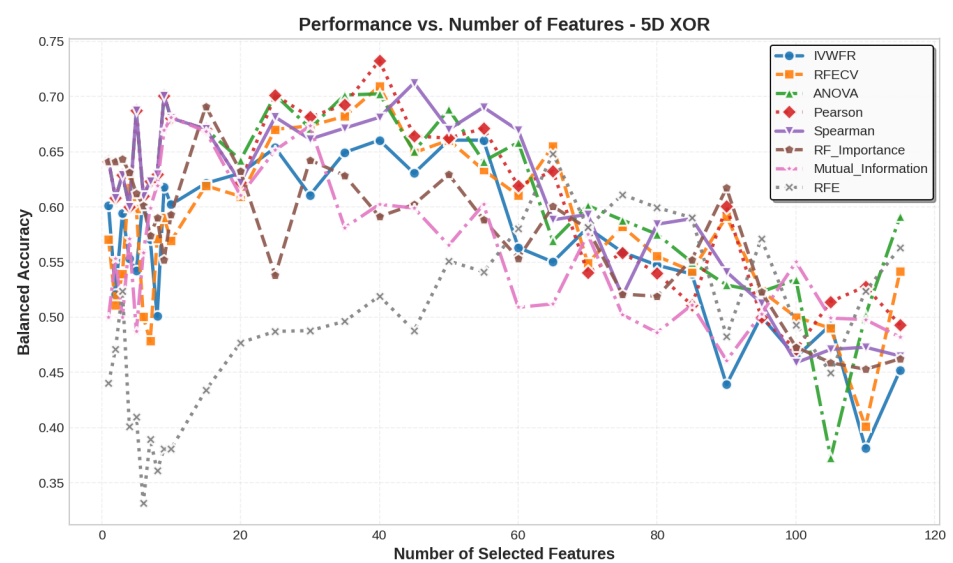


Figure A20. Performance compared to number of features using 5D XOR dataset.

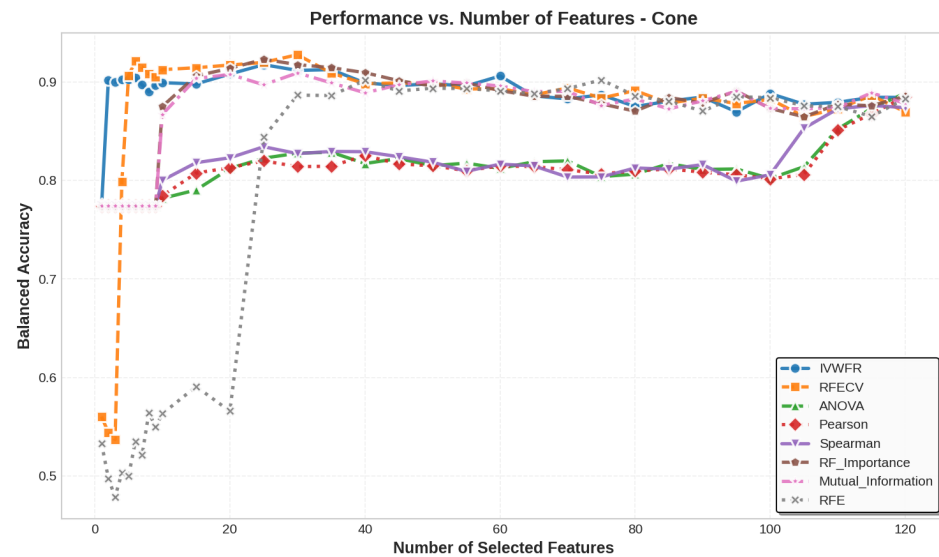


Figure A21. Performance compared to number of features using Cone dataset.

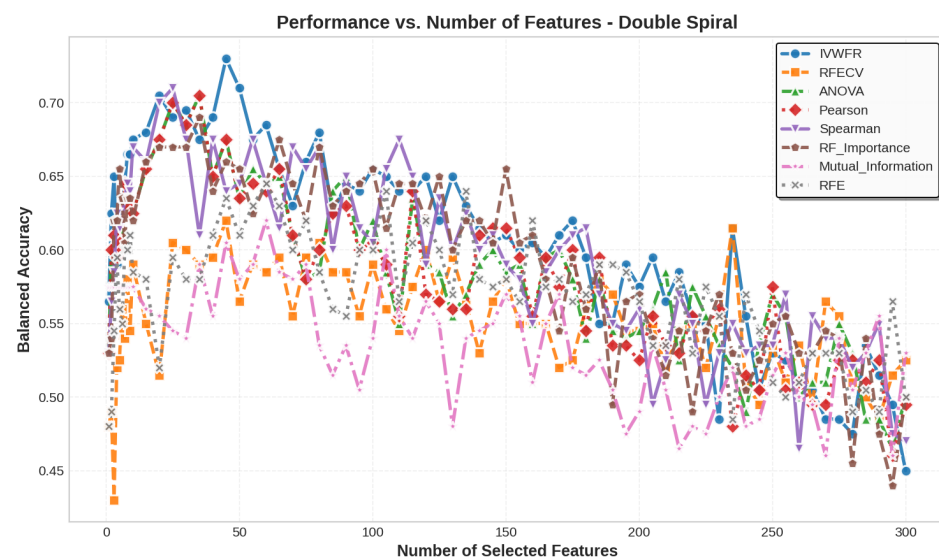


Figure A22. Performance compared to number of features using Double Spiral dataset.

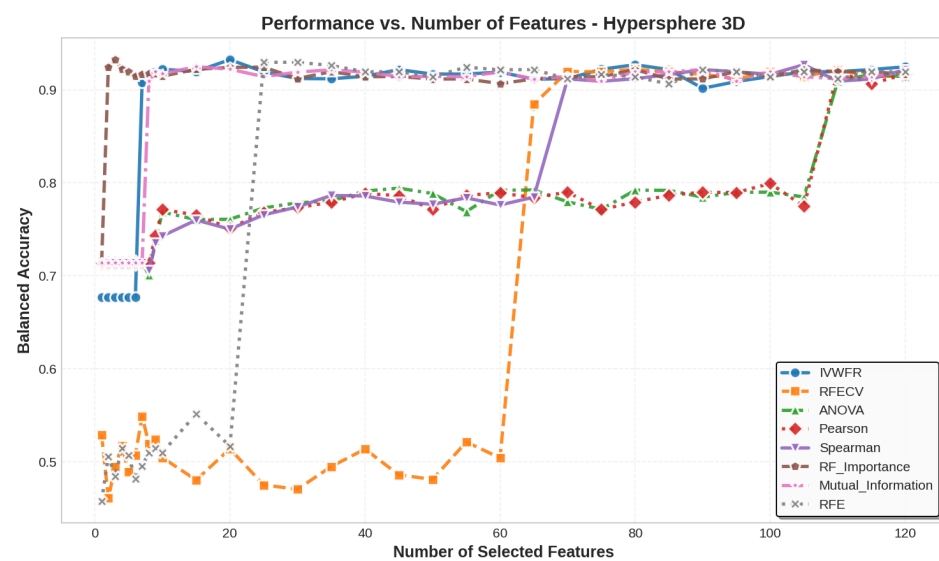


Figure A23. Performance compared to number of features using Hypersphere 3D dataset.

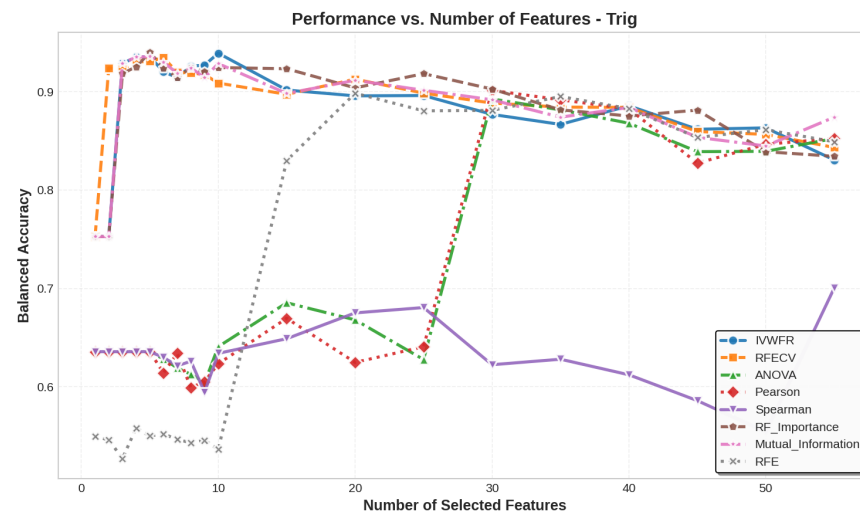


Figure A24. Performance compared to number of features using Trig dataset.

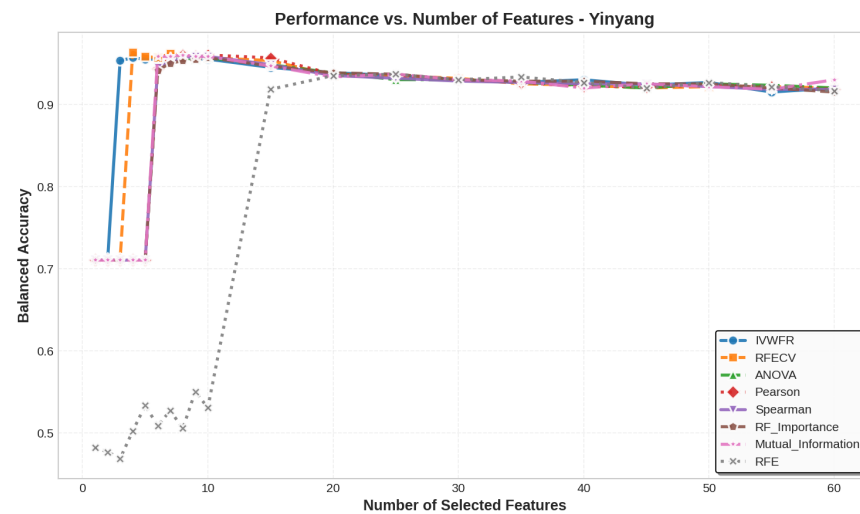


Figure A25. Performance compared to number of features using Yinyang dataset. The results for the Pearson, Spearman, RF_Importance and Mutual_Information methods and for the IVWFR and ANOVA methods are almost identical and overlap.

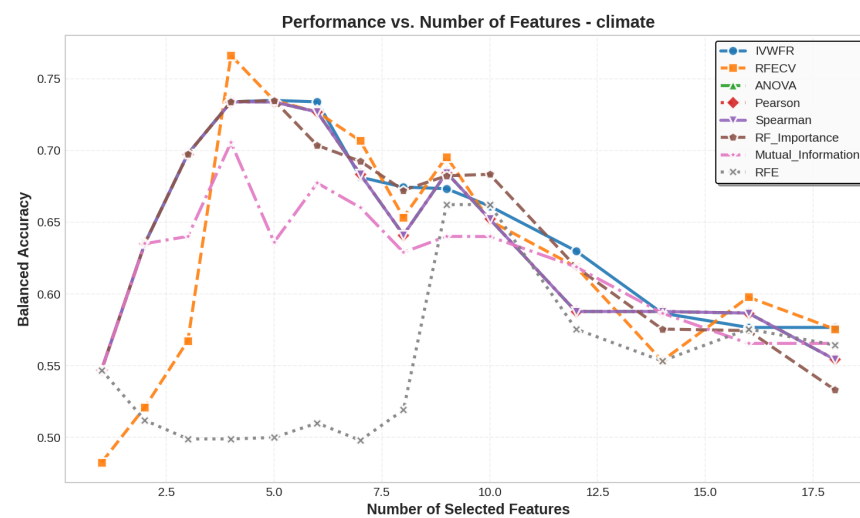


Figure A26. Performance compared to number of features using Climate dataset. The results for the ANOVA and IVWFR methods and for the Pearson and Spearman correlation methods are identical and overlap.

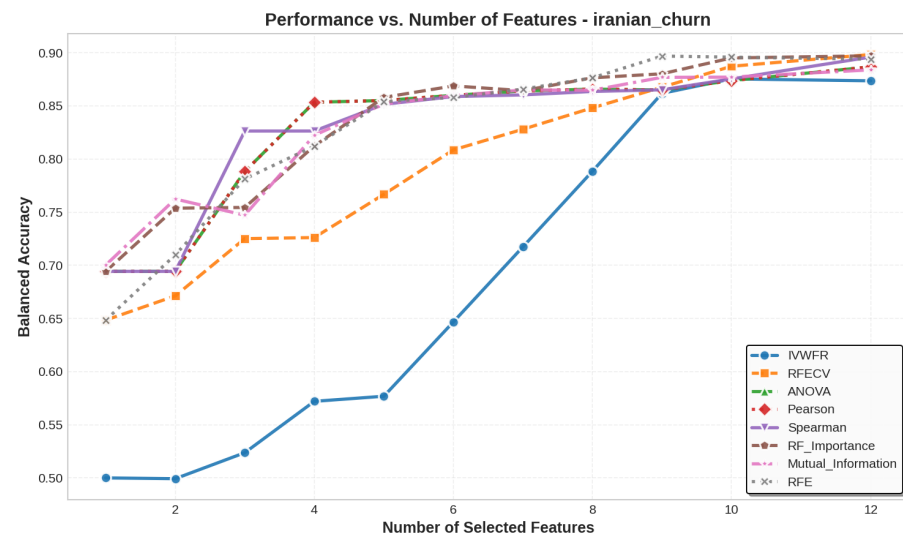


Figure A27. Performance compared to number of features using Iranian Churn dataset. The results for the ANOVA and Pearson correlation methods are identical and overlap.

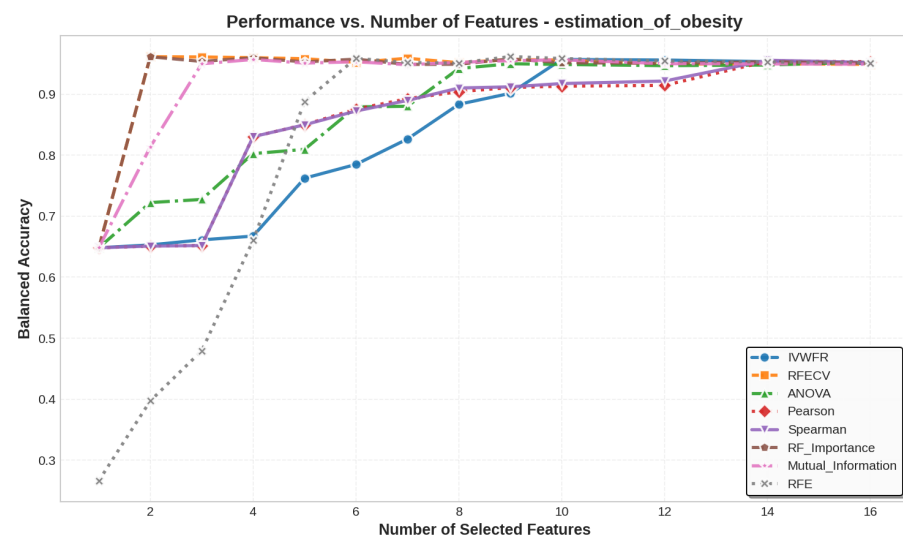


Figure A28. Performance compared to number of features using Estimation of Obesity dataset.

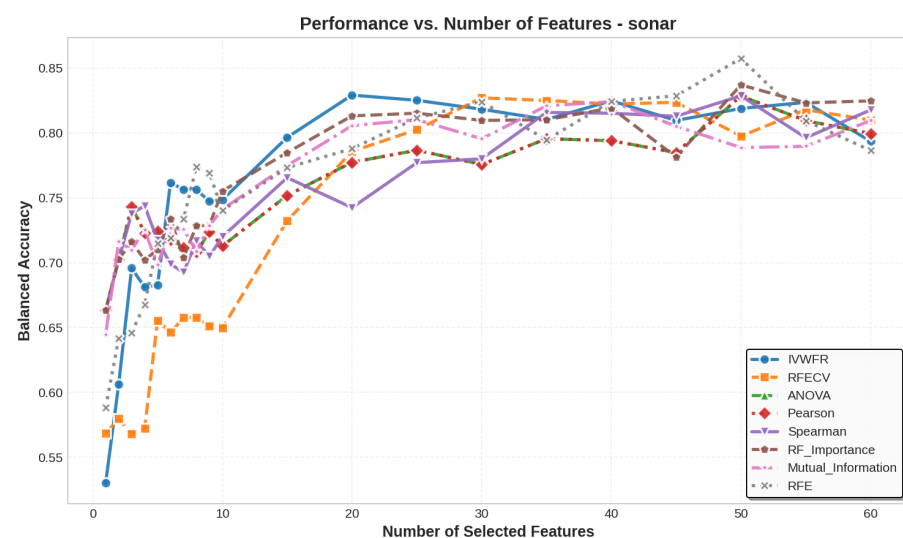


Figure A29. Performance compared to number of features using Sonar dataset. The results for the ANOVA and Pearson correlation methods are identical and overlap.

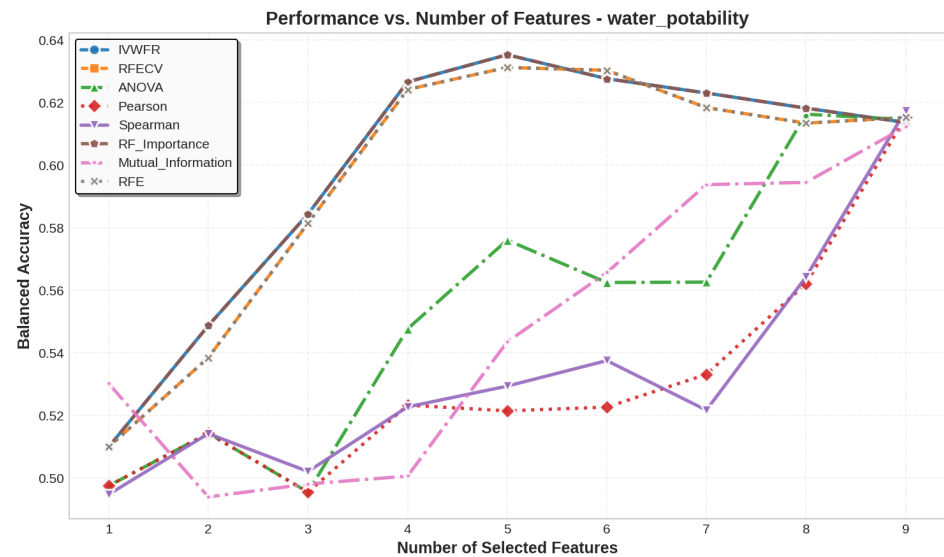


Figure A30. Performance compared to number of features using Water potability dataset. The results for the RF_Importance and IVWFR methods and for the RFE and RFECV methods are identical and overlap.

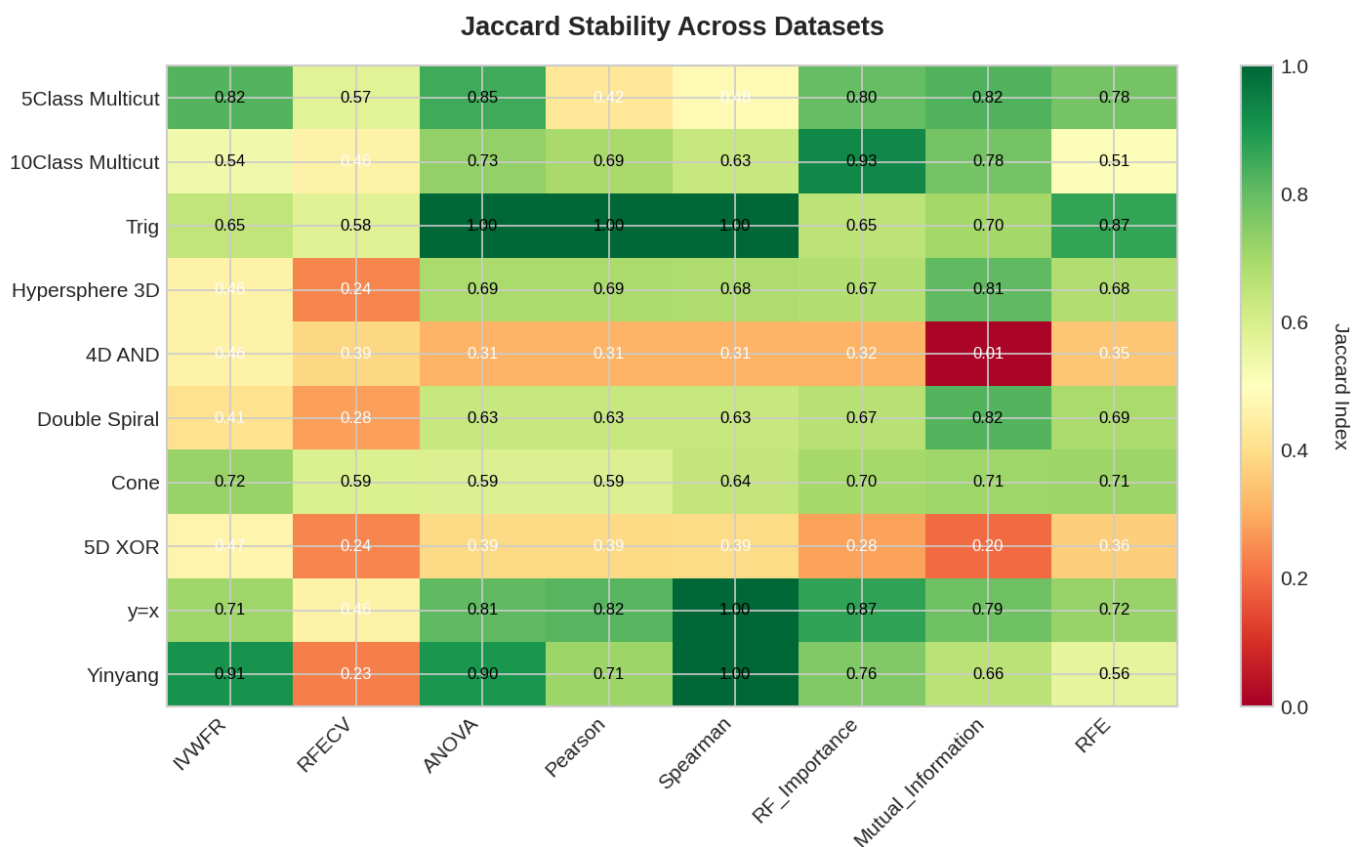


Figure A31. Jaccard stability across synthetic datasets.

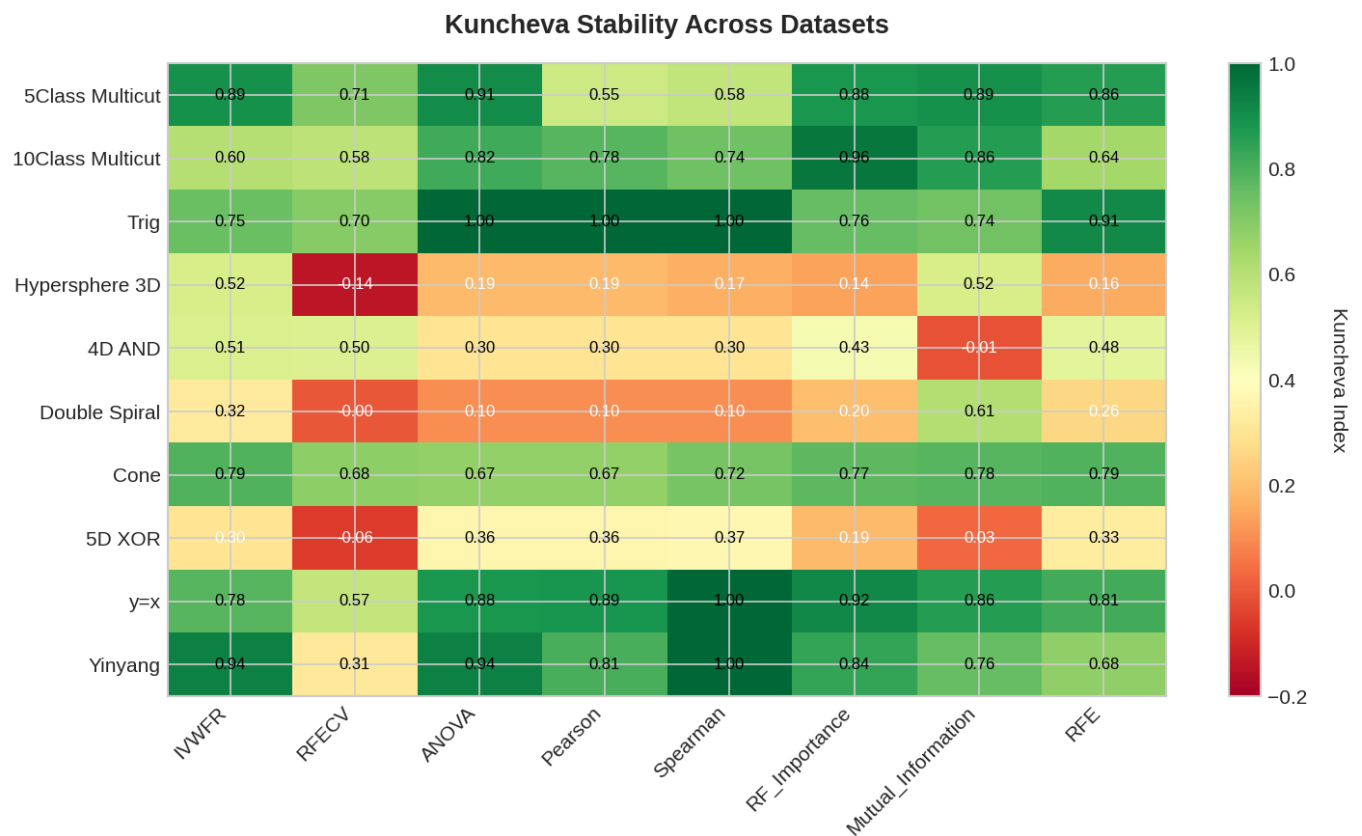


Figure A32. Kuncheva stability across synthetic datasets.

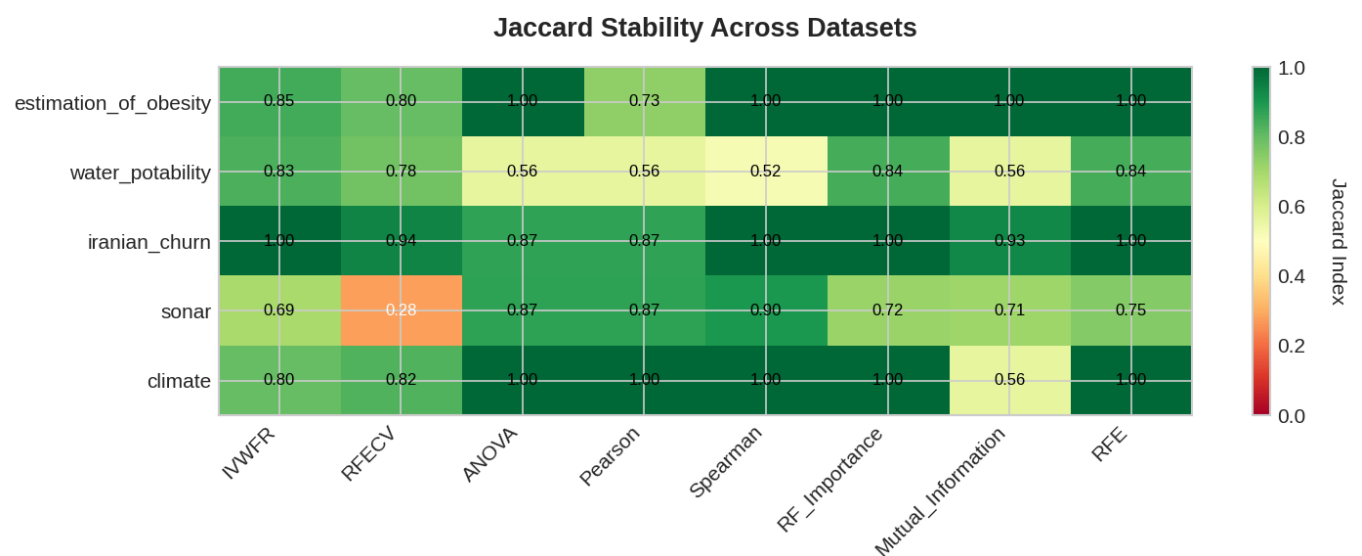


Figure A33. Jaccard stability across real-world datasets.

Table A1. Results of performance with artificial missing data for the Estimation of Obesity dataset.

Missing %	Mean Acc	Std Dev	Degradation	Avg Features
5%	0.9178	0.0049	0.0376 (3.9%)	8.0
10%	0.8912	0.0032	0.0643 (6.7%)	8.0
15%	0.8635	0.0032	0.0920 (9.6%)	8.0
20%	0.8352	0.0070	0.1203 (12.6%)	8.0

Table A2. Results of performance with artificial missing data for the Water Potability dataset.

Missing %	Mean Acc	Std Dev	Degradation	Avg Features
5%	0.6172	0.0055	0.0021 (0.3%)	5.0
10%	0.6113	0.0084	0.0080 (1.3%)	5.0
15%	0.5923	0.0087	0.0271 (4.4%)	5.0
20%	0.5948	0.0052	0.0246 (4.0%)	5.0

Table A3. Results of performance with artificial missing data for the Iranian Churn dataset.

Missing %	Mean Acc	Std Dev	Degradation	Avg Features
5%	0.8377	0.0156	0.0411 (4.7%)	7.0
10%	0.7991	0.0171	0.0797 (9.1%)	7.0
15%	0.7858	0.0046	0.0930 (10.6%)	7.0
20%	0.7724	0.0166	0.1064 (12.1%)	7.0

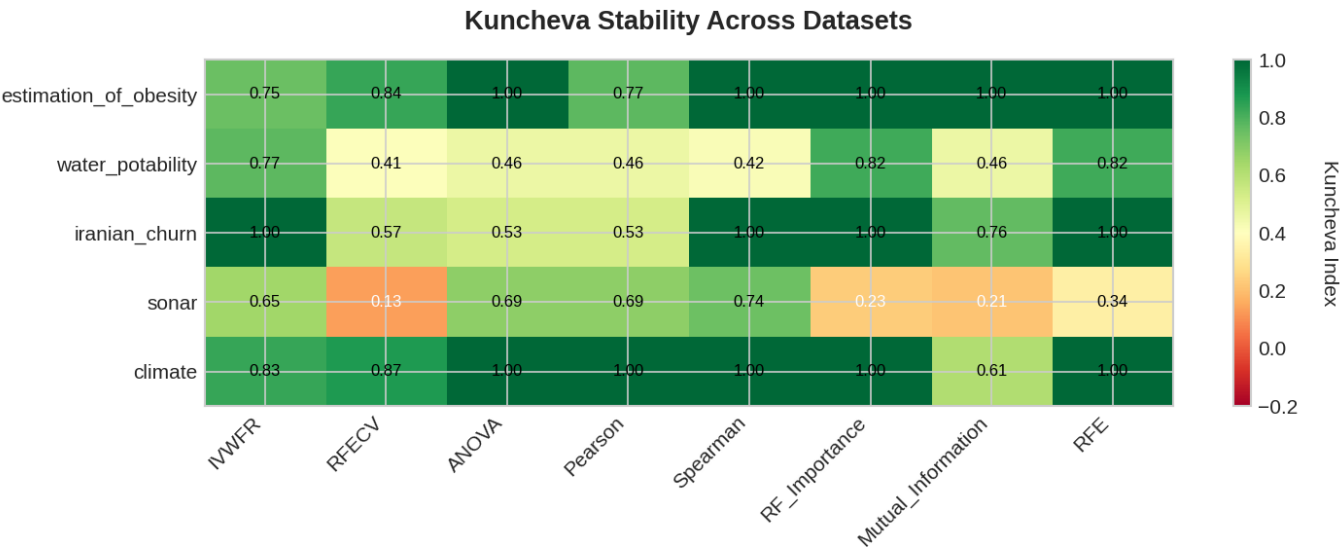


Figure A34. Kuncheva stability across real-world datasets.

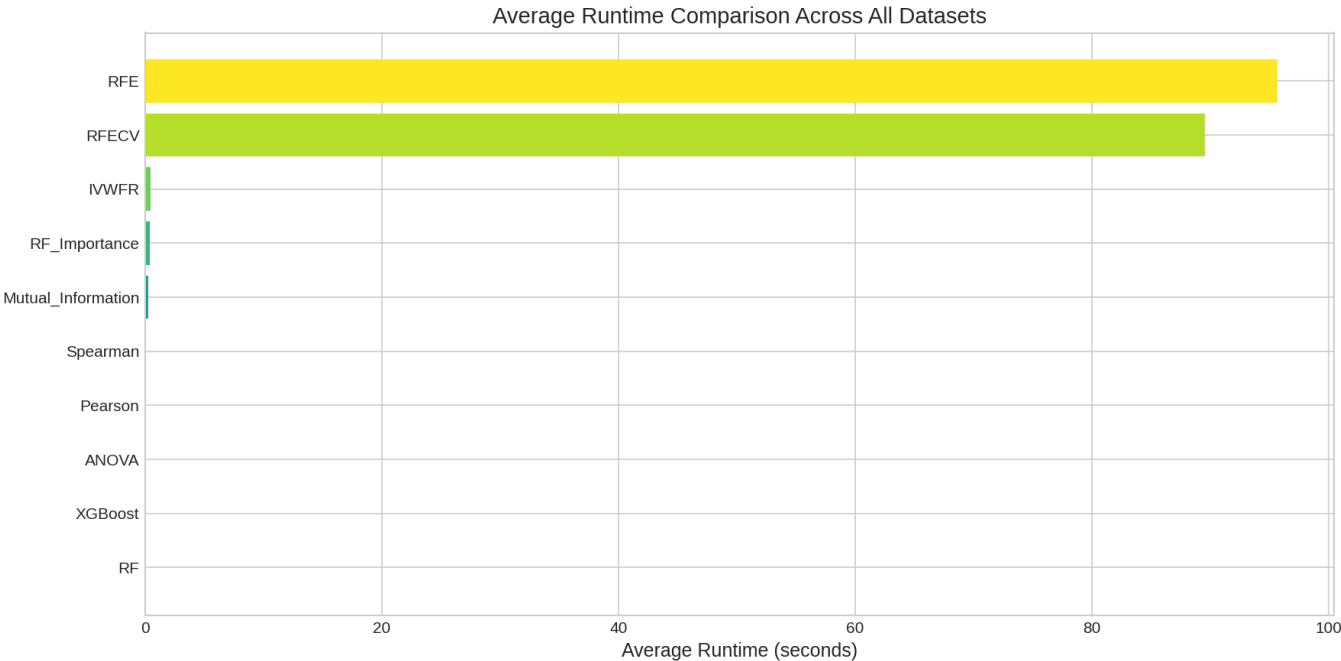


Figure A35. Average runtime across synthetic datasets.

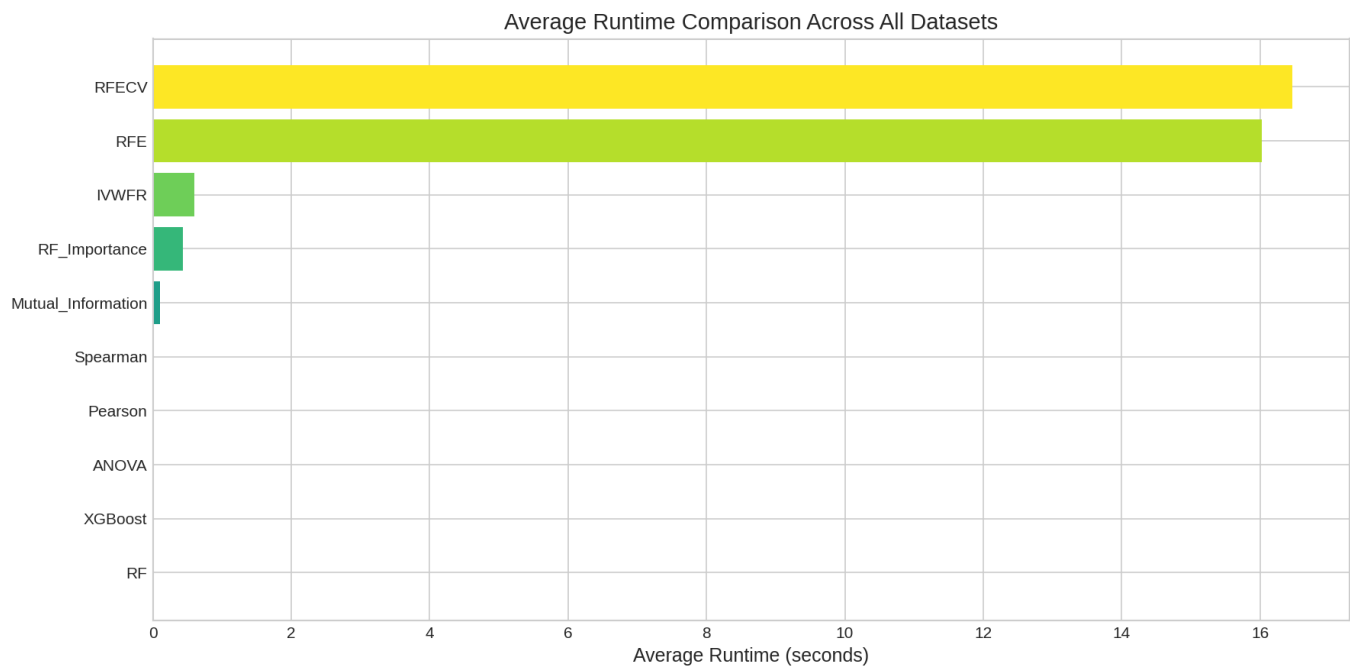


Figure A36. Average runtime across real-world datasets.

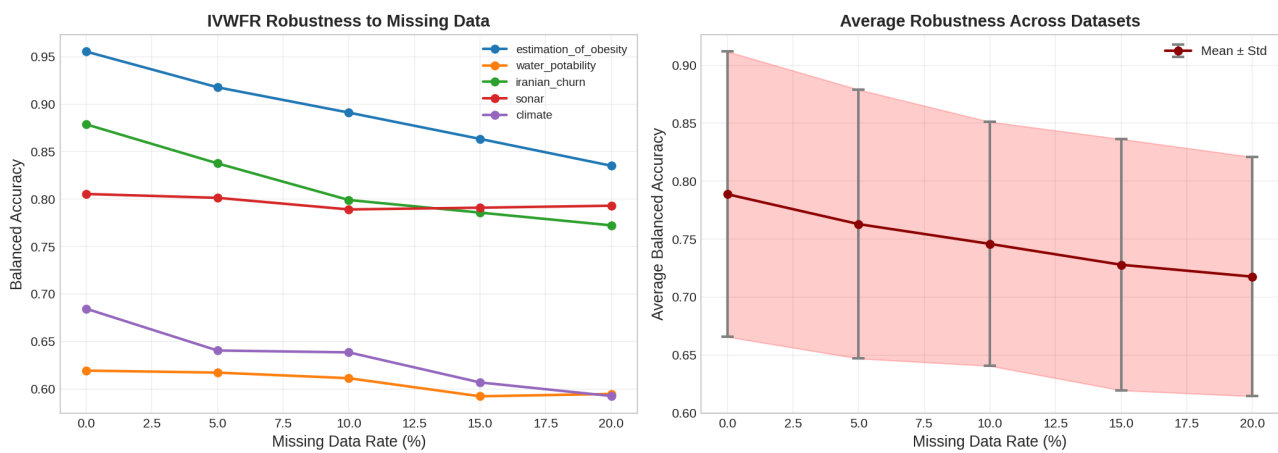


Figure A37. IVWFR algorithm robustness to missing data.

Table A4. Results of performance with artificial missing data for the Sonar dataset.

Missing %	Mean Acc	Std Dev	Degradation	Avg Features
5%	0.8014	0.0123	0.0041 (0.5%)	30.0
10%	0.7891	0.0207	0.0163 (2.0%)	30.0
15%	0.7910	0.0190	0.0144 (1.8%)	30.0
20%	0.7931	0.0219	0.0124 (1.5%)	30.0

Table A5. Results of performance with artificial missing data for the Climate dataset.

Missing %	Mean Acc	Std Dev	Degradation	Avg Features
5%	0.6405	0.0181	0.0440 (6.4%)	9.0
10%	0.6386	0.0365	0.0460 (6.7%)	9.0
15%	0.6069	0.0298	0.0777 (11.3%)	9.0
20%	0.5926	0.0232	0.0920 (13.4%)	9.0

References

1. Dash, M.; Liu, H. Feature selection for classification. *Intell. Data Anal.* **1997**, *1*, 131–156. [\[CrossRef\]](#)
2. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
3. Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [\[CrossRef\]](#)
5. Kuncheva, L.I. A stability index for feature selection. In Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and Applications, Innsbruck, Austria, 12–14 February 2007; pp. 390–395.
6. Hauray, A.C.; Gestraud, P.; Vert, J.P. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE* **2011**, *6*, e28210. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Nogueira, S.; Sechidis, K.; Brown, G. On the stability of feature selection algorithms. *J. Mach. Learn. Res.* **2018**, *18*, 1–54.
8. Cawley, G.C.; Talbot, N.L. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **2010**, *11*, 2079–2107.
9. Ahmad, J.; Latif, S.; Khan, I.U.; Alshehri, M.S.; Khan, M.S.; Alasbali, N.; Jiang, W. An interpretable deep learning framework for intrusion detection in industrial Internet of Things. *Internet Things* **2025**, *33*, 101681. [\[CrossRef\]](#)
10. Strobl, C.; Boulesteix, A.L.; Kneib, T.; Augustin, T.; Zeileis, A. Conditional variable importance for random forests. *BMC Bioinform.* **2008**, *9*, 307. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Meinshausen, N.; Bühlmann, P. Stability selection. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2010**, *72*, 417–473. [\[CrossRef\]](#)
12. Kalousis, A.; Prados, J.; Hilario, M. Stability of feature selection algorithms: A study on high-dimensional spaces. *Knowl. Inf. Syst.* **2007**, *12*, 95–116. [\[CrossRef\]](#)
13. Wojtowicz, A.; Bentkowska, U.; Paja, W. The interval-valued weighted feature ranking method based on aggregation functions. In Proceedings of the 2025 IEEE International Conference on Fuzzy Systems (FUZZ), Reims, France, 6–10 July 2025; pp. 1–7. [\[CrossRef\]](#)
14. Zadeh, L.A. The concept of a linguistic variable and its application to approximate reasoning—I. *Inf. Sci.* **1975**, *8*, 199–249. [\[CrossRef\]](#)
15. Dubois, D.; Prade, H. Internal-valued fuzzy sets, possibility theory and imprecise probability. In Proceedings of the 4th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2005) and 11èmes Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2005), Barcelona, Spain, 7–9 September 2005; pp. 314–319.
16. Komorníková, M.; Mesiar, R. Aggregation functions on bounded partially ordered sets and their classification. *Fuzzy Sets Syst.* **2011**, *175*, 48–56. [\[CrossRef\]](#)
17. Bustince, H.; Fernandez, J.; Kolesárová, A.; Mesiar, R. Generation of linear orders for intervals by means of aggregation functions. *Fuzzy Sets Syst.* **2013**, *220*, 69–77. [\[CrossRef\]](#)
18. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2623–2631. [\[CrossRef\]](#)
19. Reference Page for the Accuracy Score Implementation in Scikit-Learn. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html#sklearn.metrics.accuracy_score (accessed on 13 June 2023).
20. Reference Page for the Recall Score Implementation in Scikit-Learn. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html (accessed on 13 June 2023).
21. Reference Page for the Precision Score Implementation in Scikit-Learn. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html (accessed on 13 June 2023).
22. Reference Page for the F1 Score Implementation in Scikit-Learn. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html (accessed on 13 June 2023).
23. Mitra, R.; Ali, E.; Varam, D.; Sulieman, H.; Kamalov, F. Variable Selection in Data Analysis: A Synthetic Data Toolkit. *Mathematics* **2024**, *12*, 570. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.