IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 31 October 2023, accepted 3 December 2023, date of publication 12 December 2023,
date of current version 19 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3342044

RESEARCH ARTICLE

A More Flexible and Robust Feature Selection Algorithm

TIANYI TU[1], YE SU[1], YAYUAN TANG[2,3], WENXUE TAN[1], AND SHENG REN[1]
[1]School of Computer and Electrical Engineering, Hunan University of Arts and Science, Changde 415000, China
[2]School of Information Engineering, Hunan University of Science and Engineering, Yongzhou 425199, China
[3]School of Computer Science and Engineering, Central South University, Changsha 430083, China

Corresponding author: Tianyi Tu (tutianyi@huas.edu.cn)

This work was supported in part by the Natural Science Foundation of China under Project 62102147; in part by the National Science Foundation of Hunan Province under Project 2022JJ30424, Project 2022JJ50253, and Project 2022JJ30275; in part by the Scientific Research Project of Hunan Provincial Department of Education under Project 21B0616 and Project 21B0738; in part by the Hunan University of Arts and Sciences Ph.D. Start-Up Project under Project BSQD02; and in part by the Construct Program of Applied Characteristic Discipline at the Hunan University of Science and Engineering.

ABSTRACT With the increasing amount of real data, the challenges of large-scale model operations as well as poor generalization capacity, making selection of an appropriate feature set a significant concern. This study proposes ImprovedRFECV, an enhanced approach for cross-validated recursive feature elimination (RFECV). The algorithm first enhances the robustness of the optimal feature subset through random sampling of different data, building multiple models, and comparing their scores. Simultaneously, the L1 and L2 regularization terms are introduced to evaluate the value of each feature more comprehensively, thus reducing the impact on the anti-interference term and further improving the accuracy of the algorithm and its stability. Furthermore, a multi-model ensemble learning framework is employed to enhance generalization ability and effectively prevent overfitting. Lastly, a both-end expansion removal strategy is adopted to address the issue of strong covariance among features while enhancing the algorithm's flexibility. The experimental results demonstrate that, compared to the RFECV algorithm, the ImprovedRFECV algorithm achieves fewer optimal average features and outperforms the optimal feature subset across five datasets spanning five different domains, demonstrating the algorithm's high level of robustness and generalization ability.

INDEX TERMS Feature selection, machine learning, stability evaluation, regularization, RFECV.

I. INTRODUCTION

Feature selection [1], [2] is a common preprocessing technique in machine learning that aims to improve both the performance and interpretability of a model. It aims to identify the most optimal subset of features from the original data. Practical applications often face challenges such as the curse of dimensionality, where the dependence between features in the dataset can result in issues like increased model computation and reduced generalizability. Hence, the selection of an appropriate feature set is crucial in machine learning tasks. The field of selection employs various algorithms and techniques to tackle the aforementioned issues. These methods aim to optimize the selection process and identify the best feature subset for a given problem.

Three categories of feature selection algorithms can be distinguished based on various feature selection strategies: Filter [3], [4], [5], [6], [7], [8], Wrapper [9], [10], [11], [12], [13], [14], [15], and Embedded [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26]. The Filter method primarily relies on statistical metrics or information-theoretic measures to assess the usefulness or duplication of data aspects and thereafter choose the most advantageous variables for modeling. Liu et al. [3] designed a novel unsupervised feature selection that considers complete sample correlations and feature dependencies in a unified framework. Specifically, the significant neighbors of each sample are comprehensively retained and selected through self-representation dependency

The associate editor coordinating the review of this manuscript and approving it for publication was Christos Anagnostopoulos.

and graph construction. Additionally, mutual information is used to consider correlations between features, and sparse learning is to remove informative features. However, the Filter method, while simple to use and computationally fast, lacks control over the number of features and may select an inadequate amount, resulting in overfitting or underfitting [4]. The Wrapper method determines the quality of a feature subset by iteratively selecting candidate features, training the corresponding learner, and continuously optimizing until the optimal set of features is found. This method takes into account the interaction between features and can find a better subset of features. However, this type of algorithm, relying on training samples for modeling, is prone to overfitting. Overfitting can occur if the subset is too large or the number of training samples is too small, adversely impacting the effectiveness of feature selection [4]. Embedded [16], [17] methods utilize the feature selection capabilities embedded in specific learners, such as Lasso, Ridge, etc. This method also considers feature interactions, while not requiring additional computation, and thus is computationally less expensive. Zhang et al. [18] investigated a new feature selection method for data classification that effectively combines the discriminative power of features with ridge regression. It first establishes the global structure of the training data through linear discriminant analysis to help identify discriminative features. Then, the ridge regression model is utilized to evaluate the feature representation and discriminative information to obtain a representative coefficient matrix. The importance of the features can be calculated this representative coefficient matrix. Finally, the selected new feature subset is applied to a linear support vector machine for data classification. The method achieves good results in terms of computational efficiency and cost. While deep learning methods fall under the Embedded methods category, it is worth noting that Embedded methods rely on specific learning algorithms and may be susceptible to overfitting. If neural networks are used for feature selection, the interpretability is poor [19]. Additionally, there exist other feature selection algorithms and methods, such as the feature selection method for the weighted Gini index (WGI) proposed by Liu et al. [20]. The comparison results of feature selection methods, including Chi2, F-statistics, and Gini index, indicate that F-statistics and Chi2 exhibit superior performance when only a few features are selected. The embedded feature selection method for the weighted Gini index (WGI) has the highest probability of achieving optimal performance as the number of selected features increases. Hu R [21] et al. have combined feature selection, low-rank selection, and subspace learning into a cohesive framework. In particular, they utilize the low-rank constraint for feature selection within the context of a linear regression model. This approach carefully considers the dual aspects of information present in the data. The low-rank constraint considers the correlation among response variables and embeds an L2P-norm regularization to account for the correlation among class indicators, feature vectors, and their corresponding response variables. Additionally, they utilize the LDA algorithm, which belongs to subspace learning, to further adjust relevant feature selection results. Lastly, they conducted experiments on multiple real-life multi-view image sets. The findings confirmed that the proposed methodology outperformed all comparison algorithms. Further examples are the classical PCA [22], [23], ICA [24], [25], LapSVM [26], etc. Each of these methods has its own characteristics and application range, and it is necessary to choose the appropriate method according to the specific problem.

Currently, these feature selection algorithms have been effectively applied in various fields, especially the RFECV algorithm has been widely used in several fields. In the geological sciences field [27], it has been used for feature selection in soil heavy metal contamination data. In the electric load field [28], [29], it has been used to select. the most relevant variables to predict load demand. In the medical health domain [30], [31], [32], [33], [34], it has been used to select important features that exhibit a high correlation with certain diseases or clinical indicators. In the environmental science domain [35], it can be used to select key features related to environmental pollution and other environmental problems. In the finance field [36], it has been used for selecting features highly correlated with stock prices and market trends. However, the best subset of features selected by this algorithm is less robust and does not solve the covariance problem [37], while it is prone to overfitting [38]. Thus, there is a need to enhance existing feature selection methods to attain superior performance and overcome these limitations.

This paper presents an enhanced version of the RFECV algorithm proposed, as depicted in Figure 1. The ImprovedRFECV algorithm exhibits several advantages over the RFECV algorithm in three aspects. First, the incorporation of stability evaluation, L1 regularization term, and L2 regularization term in the ImprovedRFECV algorithm enhances the robustness of the best feature subset and mitigates the covariance issue. Moreover, the ImprovedRFECV algorithm employs an ensemble learning framework, rather than a single model, thereby enhancing the model's generalization ability and mitigating the risk of overfitting. Thirdly, the utilization of a both-end expansion removal strategy optimizes the selection step and removal strategy. This approach not only accelerates feature selection but also reduces the final selection of feature subsets, resulting in a better feature subset for the final selection. Overall, these improvements make the ImprovedRFECV algorithm as a highly effective and efficient feature selection method, surpassing the traditional RFECV algorithm, with versatile applications across various fields.

This study used five datasets from the Tianqihoubao and Kaggle platforms. The datasets consisted of air quality PM2.5 data, bike-sharing demand data, house price data, concrete strength data, and electricity load data. The original datasets, which were filtered by both the RFECV algorithm and ImprovedRFECV algorithm, were predicted using Random Forest (RF), LightGBM (LGBM), XGBoost (XGB), Gradient Boosting Decision Tree (GBDT), and

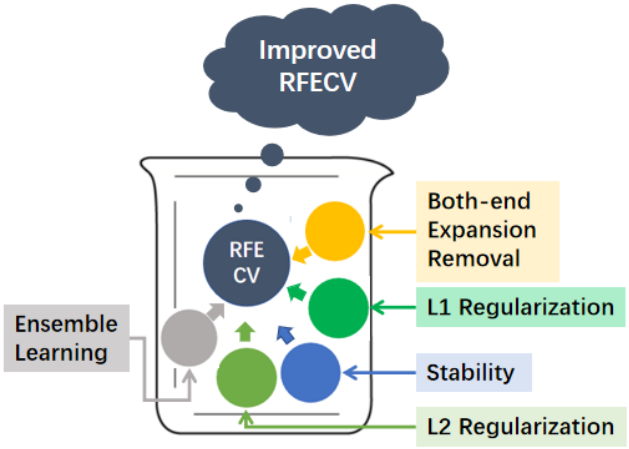

FIGURE 1. Overview figure of the ImprovedRFECV algorithm.

four ensemble tree models, respectively, using Root Mean Square Error (RMSE), Root Mean Squared Logarithmic Error (RMSLE), R-Squared (R2), and Mean Absolute Percentage Error (MAPE) to compare the performance of the original feature set, the best feature subset selected by the RFECV algorithm, and the best feature subset selected by the ImprovedRFECV algorithm. The experimental results show that the best feature subset selected by the ImprovedRFECV algorithm achieves superior performance while utilizing fewer features. This suggests that the ImprovedRFECV algorithm effectively enhances the feature selection process and improves the model's performance.

In summary, the ImprovedRFECV algorithm has several key contributions in the field of feature selection. These contributions encompass:

1) mitigating the risk of overfitting the feature selection algorithm.

2) effectively addressing the covariance problem.

3) providing control over the number of features, increasing the algorithm's flexibility.

4) accelerating feature selection and reducing computational effort.

5) improving the robustness of the optimal feature subset.

Consequently, the ImprovedRFECV algorithm is a more effective and efficient feature selection method, with applications across a broad range of fields.

II. RELATED JOB

At present, among the three types of feature selection algorithms, Filter, Wrapper, and Embedded, the wrapped feature selection algorithm is widely used, especially the RFECV algorithm has become a hot application in the field of feature selection. In the field of geoscience, for the problem of soil heavy metal pollution, Wang et al. [27] used support vector machine recursive feature elimination cross-validation (SVM-RFECV) to select among pre-selected feature bands. In the field of electric load, Liang et al. [28] proposed a two-stage short-term load forecasting method based on the RFECV algorithm and the time-convolutional network efficient channel attention mechanism-long and short-term memory network (TCN-ECA-LSTM). Veljanovski et al. [29] proposed a method to combine the ability of neural networks to learn nonlinear relationships between features with the optimization capability of the RFECV algorithm to find the best prediction model. In the field of medical health, Sung et al. [30] proposed a new stroke severity classification method that uses symmetric gait features and the RFECV algorithm, and experiments showed that combining symmetric gait data with the RFECV technique can improve the classification performance of stroke severity. Ossai et al. [31] developed a method based on cross-validation and additional tree classifier recursive feature elimination machine learning algorithm (RFECV-ETC) to predict extended pre-hospital stay (ELOHS) and its risk factors with very good results. Amakrane et al. [32] proposed a new handwritten feature selection method to obtain relevant features to effectively identify Parkinson's disease. The method is based on the REFCV algorithm to determine the best classifier for predicting Parkinson's disease. Hou et al. [33] proposed an SVM-RFECV algorithm for predicting Alzheimer's disease. Assegie et al. [34] used the RFECV algorithm to explore the effect of heart disease feature quality on the prediction performance of machine learning models for heart disease. In environmental science, Tong et al. [35] used the RFECV algorithm, random forest feature selector (RFFS), and principal component analysis (PCA) to optimize sensor arrays for real-time and fast detection of automobile exhaust pollutants. In the field of finance, Bamunuarachchi and Silva [36] used a logistic regression RFECV feature elimination model to evaluate the performance of pawn operations.

In conclusion, the RFECV algorithm can improve the prediction accuracy of the model while reducing computational time and resource consumption. Moreover, the RFECV algorithm ensures the robustness of the model by nested cross-validation and prevents the bias of the results due to sample reassignment.

III. MOTIVATION

Although widely used in various fields, the traditional RFECV algorithm still has limitations.

1) The accuracy and stability of traditional RFECV algorithms are affected by the anti-interference term. For instance, when there is high covariance or the presence of outliers, RFECV algorithms may incorrectly select less useful features, thereby weakening the model's predictive power [37].
2) The traditional RFECV algorithm's lack of flexibility comes from removing one feature in each iteration round.
3) The performance and stability of traditional RFECV algorithms are dependent on the chosen classifier or regression method. Different methods can yield varying feature subsets, which can impact the model's performance and stability [38].

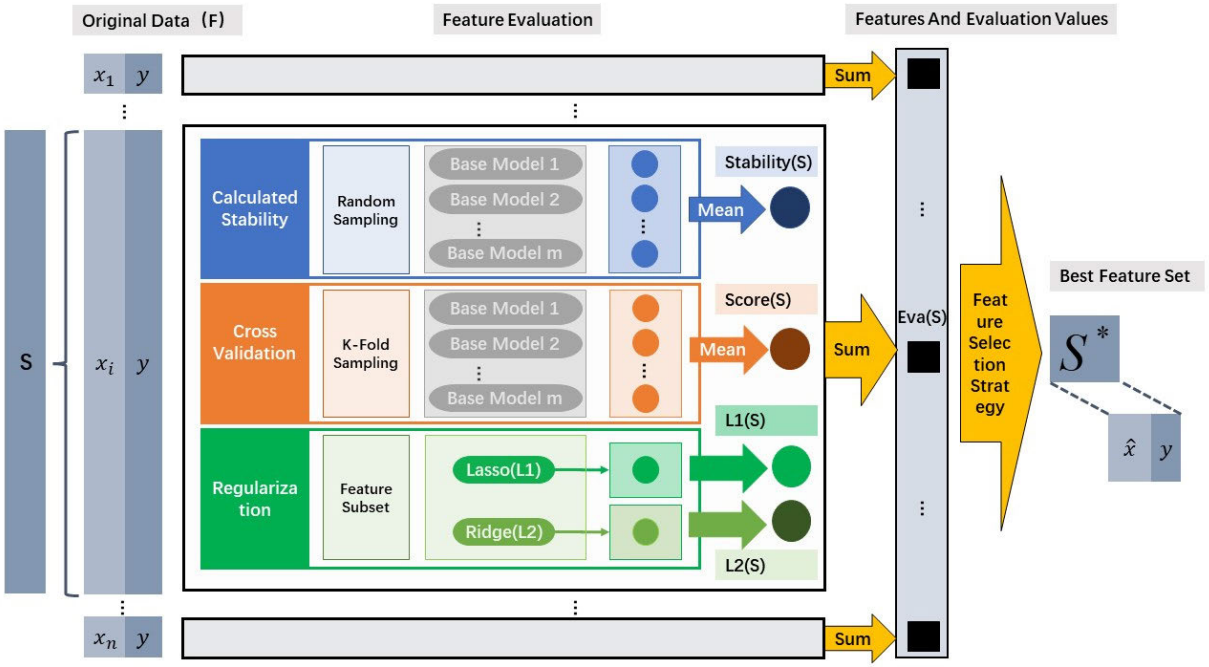

FIGURE 2. ImprovedRFECV algorithm schematic.

4) The traditional RFECV algorithm's lack of stability evaluation methods can result in uncertainty and bias when dealing with random sample sampling and perturbation. As a consequence, the algorithm's results may be influenced by specific datasets and cross-validation collapses. This can lead to the selection of a suboptimal feature subset that lacks robustness and generalizability to other datasets.

Based on the above-mentioned shortcomings of the traditional RFECV algorithm, we propose an ImprovedRFECV algorithm. The improvement of this algorithm for the above four drawbacks is as follows.

To address the drawbacks 1), the regularization term is used to alleviate the covariance problem. When there are redundant features in the dataset, L1 regularization tends to select one of the features and ignore other highly correlated features, thus avoiding the covariance problem. In contrast, L2 regularization penalizes the sum of squares of the regression coefficients by adding an L2 parametric term to the optimization objective function. Unlike L1 regularization, L2 regularization will make the regression coefficients as close to 0 as possible but does not directly compress them to 0. Moreover, L2 regularization will make the weights of multiple highly correlated features similar for them.

To address the drawbacks 2), the number of features selected in each round and the step size can be set according to the actual problem, to better control the flexibility and effectiveness of the algorithm. Also use a strategy of expanding removal at both ends to speed up feature selection, reduce computation, and further improve algorithm efficiency.

To address the drawback 3), The L1 regularization and L2 regularization terms are introduced to calculate the scores of the features, and integrated learning is used to prevent overfitting, thus ensuring the performance and stability of the model. the L1 regularization term is used to help select between features with equal contributions and some irrelevant feature coefficients can be set to 0. the L2 regularization term is used to prevent the feature coefficients from being too large and reduce the complexity of the model.

To address the drawback 4), by adding a stability assessment method to each regression model and considering both model performance and stability requirements when selecting features. In each iteration, we randomly sampled a portion of the training data and used this portion to train the model and calculate the average score of the model. Since we may randomly sample multiple times, the stability of a subset of features can be estimated in this way, thus increasing their robustness.

IV. METHOD

The principle of the ImprovedRFECV algorithm consists of two parts. The first part involves feature evaluation, while the second part focuses on the feature selection strategy. Figure 2 illustrates the specific principles of the algorithm.

A. FEATURE EVALUATION

Let the initialized feature set be F, and select the feature subset S, namely $S \subseteq F$, and the evaluation value of the feature subset shall be:

$$\text{Eva}(S) = \text{Score}(S) + \text{Stability}(S) + L_1(S) + L_2(S) \quad (1)$$

where Score (S) is the score obtained by the cross-validation method, Stability(S) is the score obtained by using the random sampling method, L_1 (S) is the score of the L1 regularization term, and $L_2(S)$ is the score of the L2 regularization term.

1)Score

In the process of cross-validation, the data is split into k subsets or folds, with one-fold selected as the validation set and the remaining k-1 folds as the training set. This process is repeated k times, each time with a different fold designated as the validation set. The results of each model evaluation are then returned in an array. By using cross-validation, the generalization ability of a model can be more accurately evaluated. This is because, rather than evaluating the model on a single testing set, cross-validation utilizes multiple testing sets, reducing the impact of random variations and producing more reliable results. Moreover, integrating models through cross-validation further enhances the generalization ability of the model. In this approach, the k-fold cross-validation scores of each model are taken into account equally, and finally, the mean score is used as the final cross-validation score. This strategy helps to avoid overfitting and promotes the robustness of the model.

$$\text{score}(S) = \frac{1}{n} \times \frac{1}{k} \sum_{i=1}^{k} E(y_{test}^{(i)}, H(X_{train,S}^{(i)})) \tag{2}$$

where score(S) denotes the average evaluation metric value obtained from the regression model corresponding to the feature subset S under k fold-cross validation; n denotes the number of models; $X_{train,S}^{(i)}$ denotes that only the features in S are selected as input features in the training set of the ith validation set; $y_{test}^{(i)}$ denotes the test set label of the ith validation set; $E\left(y_{test}^{(i)}, H(X_{train,S}^{(i)})\right)$ denotes the evaluation metric predicted for the ith validation set; and $H\left(X_{train,S}^{(i)}\right)$ denotes the prediction result of the corresponding regression model.

2) Stability

The stability selection method is used to assess the significance of features in a dataset by random data, constructing multiple models, and comparing scores. The specific procedure is outlined as follows:

a. Randomly select a portion of the original data to form the training and testing set.
b. Use these data to train the models and calculate their scores.
c. Use the obtained model to predict the sample labels and calculate the average difference between the predicted results of this model and the actual labels.
d. Calculate the average evaluation metric of this feature subset across all models during the sampling process as the stability score.
e. Repeat the above steps until N repetitions of sampling are completed.
f. Ultimately, evaluate the stability score for each feature subset.

A higher stability score indicates a greater contribution of the feature to the model, thereby warranting its inclusion.

$$\text{stability}(S) = \frac{1}{n_{iter}} \sum_{j=1}^{n_{iter}} E(y_{test}^{(j)}, H(X_{train,S}^{(j)})) \tag{3}$$

where stability(S) denotes the stability score corresponding to the feature subset S; n_{iter} denotes the number of samples (default is 10); $X_{train,S}^{(j)}$ and $y_{test}^{(j)}$ the training and testing sets in the jth cross-validation, respectively.

3) L1 and L2

To enhance the control of overfitting, in addition to the above stability scores, L1 and L2 regularization term scores are introduced as well. L1 regularization term penalizes the redundant or unimportant features in the feature subset, reducing noise and filtering out irrelevant ones. L2 regularization term reduces the complexity and generalization error improving prediction capability.

$$L_1(S) = \sum_{f \in S} \left| w_f^{Lasso} \right| \tag{4}$$

where $L_1(S)$ denotes the L1 regularization term score corresponding to the feature subset S; w_f^{Lasso} is the weight coefficient obtained after applying L1 regularization to the features f.

$$L_2(S) = \sum_{f \in S} w_f^{Ridge} \tag{5}$$

where $L_2(S)$ denotes the L1 regularization term score corresponding to the feature subset S; w_f^{Ridge} is the weight coefficient obtained after applying L1 regularization to the features f.

In the feature evaluation stage of the ImprovedRFECV algorithm, the scores of the current feature subset, stability score, L1 regularization term score, and L2 regularization term score are all taken into account when calculating the total score. This comprehensive approach provides a more accurate assessment of the importance of each feature in the selection process. Additionally, the ImprovedRFECV algorithm uses a model integration approach to further improve its generalization capability. This approach involves training multiple models on different subsets of the data and then combining them to produce a final prediction. By using an ensemble of models, the algorithm can better account for variations in the data and prevent overfitting, leading to more robust and reliable results.

B. FEATURE SELECTION STRATEGY

The feature selection process employs an elimination approach, which consists of the following.

1) Initialize all features by adding them to the feature set F;

$$S_0 = F, S^* = \emptyset \tag{6}$$

2) For each iteration $i = 0, 1, 2, \ldots, m$, the following operation is performed:

Algorithm 1 ImprovedRFECV

Input: feature matrix X, label vector y, number of selected features n_features, and other parameters.
Output: selected features
1: Initialize the feature set as all indices of the features.
2: Initialize an empty list selected_features to store the selected feature indices.
3: **repeat**
4: **for** each feature i in the feature set **do**
5: Initialize scores[i], stabilities[i], lasso_regularizations[i], and ridge_regularizations[i].
6: For each regression model, compute the average R-squared value of the feature subset {Xi} using cross-validation and calculate the stability score.
7: If LI regularization coefficient alpha_lasso $>$ 0, use Lasso model to compute the LI regularization score regularization_score_lasso.
8: If L2 regularization coefficient alpha_ridge $>$ 0, use Ridge model to compute the L2 regularization score regularization_score_ridge.
9: Sum up scores[i], stabilities[i], lasso_regularizations[i], and ridge_regularizations[i] to get the final score final_scores[i].
10: **end for**
11: Select the index of the feature with the highest final score, best_feature_idx = argmax(final_scores).
12: Add the feature index corresponding to best_feature_idx to selected_features.
13: Update the feature set by removing the selected feature and its step-1 adjacent features.
14: **until** n_features times
15: **return** selected features.

a. Find the optimal feature subset S for this iteration among the remaining feature subsets S_i;

$$S = \text{argmax}(Eva(S_i)) \tag{7}$$

b. Add S to the final feature subset S^*;

$$S^* = S^* + S \tag{8}$$

c. Adjustment of the remaining feature set with a both-end expansion removal method, Where S_{left} represents step-1 feature on the left of S, and S_{right} represents step-1 feature on the right of S;

$$S_{i+1} = S_i - S - S_{left} - S_{right} \tag{9}$$

d. If the number of features in S^* meets the requirement, the iteration is stopped, otherwise, the next iteration is continued;

3) Return the final feature subset, S^*.

The number of features to be skipped in each feature selection is 1. It is important to consider the boundary conditions to prevent exceeding the range of the feature index. This feature update strategy helps prevent the selection of highly correlated features, accelerates the feature selection process, and mitigates the impact of redundant features on the evaluation.

C. THE OVERALL ALGORITHM

The idea of the ImprovedRFECV algorithm is a feature screening method for model integration based on data stability, combined with L1 regularization and L2 regularization, which can effectively evaluate the impact of features on model performance and search in the feature space while using a both-end expansion removal strategy to find the optimal subset of features, thus improving the generalization capability and accuracy of the model. The implementation is shown in Algorithm 1.

V. EXPERIMENT

A. EXPERIMENTAL SETUP

To verify the performance of the proposed ImprovedRFECV algorithm, five datasets from five domains were selected for experimentation. These datasets were derived from Tianqihoubao and Kaggle platforms.

For environmental science, air quality data from the Tianqihoubao platform for Shenzhen, Guangdong Province were selected. This dataset encompasses data collected from October 31, 2013, to March 31, 2022, and primarily serves the purpose of predicting PM2.5 concentrations in the air. Datasets from the Kaggle platform were chosen for smart transportation, the financial sector, construction engineering, and electric load analysis. As an illustration, data on the count of shared bicycle rentals in Washington, DC, USA [39] was chosen primarily to forecast the demand for shared bicycles in Washington, USA. In addition, California housing price data [40] was also selected, which mainly predicts California housing prices, while concrete strength data [41] mainly predicts concrete strength and historical PDB electricity demand data mainly predicts PDB electricity demand.

The feature selection algorithm evaluates the goodness of the filtered feature subset by using the regression accuracy of the regression method. Therefore, in this study, four regression evaluation metrics, namely RMSE, RMSLE, R2, and MAPE, were used to evaluate the original dataset, the dataset screened by the RFECV algorithm, and

ImprovedRFECV algorithm, respectively, using four ensemble tree models, including RF, LGBM, XGB, and GBDT. The formulas for calculating RMSE, RMSLE, R2, and MAPE are defined as follows.

1) RMSE: RMSE is a measure of the standard deviation of the difference between the predicted and true values. It is calculated by squaring the prediction error of each sample, taking the average value, and then opening the square. the smaller the RMSE, the better the fit of the model.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{10}$$

where, n denotes the sample size, y_i denotes the true value, and $\hat{y}_i$ denotes the predicted value.

2) RMSLE: RMSLE is an error measure proposed to deal with large variations in the data. The smaller the RMSLE, the more accurate the model's prediction.

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(y_i + 1) - log(\hat{y}_i + 1))^2} \tag{11}$$

where, n denotes the sample size, y_i denotes the true value, and $\hat{y}_i$ denotes the predicted value. In RMLE, both true and predicted values are calculated by adding 1 and taking the logarithm.

3) R2: R2 is a performance metric for regression problems that measures the correlation between the predicted and true values of the model. r2 takes values between 0 and 1, with values closer to 1 indicating better prediction and closer to 0 indicating poorer prediction.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{12}$$

where, n denotes the sample size, y_i denotes the true value, and $\hat{y}_i$ denotes the predicted value, $\bar{y}_i$ indicates the average of the true values.

4) MAPE: MAPE is a measure of the magnitude of the error between the predicted value and the true value. MAPE is calculated by dividing the prediction error of each sample by the true value, and then summing and averaging the values and multiplying by 100%. a smaller MAPE indicates a smaller prediction error of the model.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\% \tag{13}$$

where, n denotes the sample size, y_i denotes the true value, and $\hat{y}_i$ denotes the predicted value.

For this experiment, we divided the entire dataset into a test set consisting of the first 20% of the data and a training set consisting of the remaining 80%. We conducted the experiments without making any optimization adjustments to the model parameters. To ensure the repeatability of the experimental results, we set the random state value of the ensemble tree models. We used the default values for all other parameters predefined in the models. The experiment was divided into two parts: the feature subset comparison experiment and the performance verification experiment.

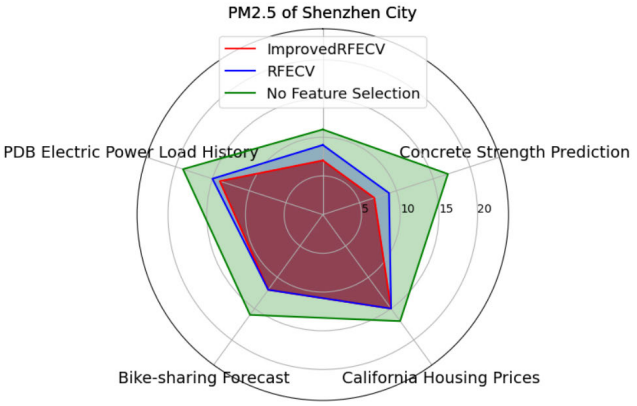

FIGURE 3. The average size of the best feature subset obtained using different methods.

B. FEATURE SUBSET COMPARISON EXPERIMENT

The feature subset comparison experiment aims to compare the optimal feature subset size for the RFECV algorithm and the ImprovedRFECV algorithm. As shown in Table 1.

In the field of environmental science, the raw data of PM2.5 air quality in Shenzhen, Guangdong Province has 3032 sample points and 10 features. In the field of finance, the raw data of house prices in California consists of 37,137 sample points and 9 features. In the field of smart transportation, the raw data of bicycle sharing rental volume in Washington, DC, USA, consists of 10,886 sample points and 12 features. In the field of construction engineering, the raw concrete strength data contains 1030 sample points and 10 features. In the field of electric load, the PDB historical electricity raw data has 103,776 sample points and 8 features.

As can be seen from Table 1, after feature engineering, the number of features in the five datasets for the five domains is 11, 17, 16, 17, and 19, respectively. The number of features decreases after applying both the RFECV algorithm and the ImprovedRFECV algorithm. The ImprovedRFECV algorithm screened the best feature subsets with an average of two fewer features compared to the RFECV algorithm in both the environmental science and construction engineering domains. The best feature subsets in the financial, electric load, and smart transportation domains were roughly the same size for both algorithms. Figure 3 further demonstrates that the ImprovedRFECV algorithm has a smaller average best feature subset size compared to the RFECV algorithm. Next, performance test verification experiments will be conducted to assess the combined performance of the two algorithms.

C. PERFORMANCE TEST VERIFICATION EXPERIMENT

In this experiment, four ensemble tree models (RF, LGBM, XGB, GBDT) were adopted for regression prediction of

TABLE 1. Scale of candidate feature subsets for the four integrated tree models on the five data sets.

Dataset	Models	Original	Feature Engineering	RFECV	Improved RFECV
PM2.5 of Shenzhen City	RF	10	11	9	7
	LGBM	10	11	10	7
	XGB	10	11	9	7
	GBDT	10	11	9	7
	Average	10	11	**9.25**	**7**
California Housing Prices	RF	9	17	15	15
	LGBM	9	17	16	15
	XGB	9	17	14	15
	GBDT	9	17	13	15
	Average	9	17	**14.5**	**15**
Bike-sharing Forecast	RF	12	16	10	12
	LGBM	12	16	15	12
	XGB	12	16	13	12
	GBDT	12	16	11	12
	Average	12	16	**12.25**	**12**
Concrete Strength Prediction	RF	10	17	15	7
	LGBM	10	17	13	7
	XGB	10	17	1	7
	GBDT	10	17	8	7
	Average	10	17	**9.25**	**7**
PDB Electric Power Load History	RF	8	19	18	14
	LGBM	8	19	15	14
	XGB	8	19	13	14
	GBDT	8	19	13	14
	Average	8	19	**14.75**	**14**

TABLE 2. Prediction accuracy of the four integrated tree models on the five data sets.

Dataset	Models	Evaluation Metrics	No feature selection	REFCV	Improved RFECV
PM2.5 of Shenzhen City	RF	R2	0.91298	0.91291	**0.91504**
	LGBM		0.9179	0.9179	**0.91935**
	XGB		0.91127	0.91039	**0.92013**
	GBDT		0.90484	0.90409	**0.90944**
California Housing Prices	RF	RMSE	**0.57401**	0.57472	0.57518
	LGBM		0.55988	0.55988	**0.55893**
	XGB		0.57624	0.57352	**0.57279**
	GBDT		0.57789	0.57852	**0.57733**
Bike-sharing Forecast	RF	RMSLE	0.13757	0.13754	**0.13512**
	LGBM		**0.10983**	**0.10983**	0.11218
	XGB		0.1155	0.11428	**0.11425**
	GBDT		0.1234	0.12396	**0.11887**
Concrete Strength Prediction	RF	RMSE	12.27744	12.28099	**12.1841**
	LGBM		11.99891	11.99501	**11.87498**
	XGB		12.47979	**12.15567**	12.24577
	GBDT		11.99425	12.01419	**11.88335**
PDB Electric Power Load History	RF	MAPE	**0.02881**	**0.02881**	0.0341
	LGBM		0.03461	0.03655	**0.03455**
	XGB		0.03306	**0.03184**	0.03574
	GBDT		0.04247	0.04232	**0.04066**
Total	-	-	3	4	**15**

problems in five different domains, and four evaluation metrics (R2, RMSE, RMSLE, MAPE) were used for validation measurement. The experimental results are shown in Table 2.

This experiment employed a range of evaluation metrics to evaluate the performance of machine learning models across different domains. In the field of environmental science, the R2 metric was chosen as it offers comprehensive insights into the quality of the fit, including the significance of the linear regression model and suitability in the data science model. This metric is especially applicable in the case of PM2.5 prediction [42]. For bike-sharing demand forecasting in the smart transportation domain, RMSLE was chosen. This metric penalizes underprediction more than overprediction, making it suitable for predicting shared bicycle demand [43]. In the financial field, a small number of outliers and outliers are usually found in the house price prediction class of problems due to the characteristics of the data set. And RMSE can handle these outliers more sensitively [44]. Therefore, RMSE is used as an evaluation metric. In the field of electric load, MAPE has less influence on extreme values (e.g., value 0) and magnitudes, making it more suitable for assessing the performance of electric load forecasting models [45]. In the field of construction engineering, RMSE was deemed appropriate as it reflects the magnitude of the error between the predicted results of the model and the actual values, which is important for concrete strength prediction models [46], [47].

As Table 2 and Figure 4 illustrate, we evaluated the regression prediction results on five datasets from five domains. We used four ensemble tree models and conducted a total of 20 measurements. The performance of the models was compared between those without feature selection and those utilizing the RFECV and ImprovedRFECV algorithms. The findings indicate that the ensemble tree models achieved the highest scores in three cases without feature selection, four cases when combined with the RFECV algorithm, and fifteen cases when combined with the ImprovedRFECV algorithm. These results demonstrate the capability of the ImprovedRFECV algorithm in conducting a more comprehensive

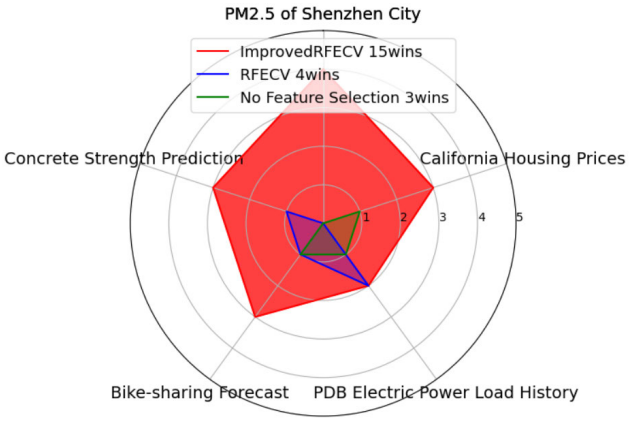

FIGURE 4. Algorithm performance comparison.

evaluation of feature subsets, leading to a more effective screening than the RFECV algorithm. This indicates that the ImprovedRFECV algorithm can enhance the accuracy and generalization capability of machine learning models. Overall, the experimental results support the reliability and effectiveness of the ImprovedRFECV algorithm in feature selection across various domains.

Analysis of Figure 5 reveals that the prediction accuracy of most of the models is reduced using the best subset of features selected by the RFECV algorithm. This is due to the presence of strongly covariant features in the dataset. RF-RFECV and XGB-RFECV solved the covariance problem of one set of features, and the prediction accuracy was improved. The prediction accuracy of both LGBM and GBDT is reduced. Although the ImprovedRFECV algorithm solves only one set of covariance features, only the prediction accuracy of LGBM has a thousand-digit decrease, while all other models have improved. As shown in Fig. 5(c), the RMSLEs of RF, XGB, and GBDT are all smaller than those of the models corresponding to the best subset of features screened using RFECV and thus have higher prediction accuracy. There are 5 sets of covariance features in the concrete strength prediction dataset. RF-RFECV solves only one set of covariance features, LGBM-RFECV solves only 2 sets of covariance features, GBDT-RFECV solves only 3 sets of covariance features, and the best feature subset of XGB-RFECV has only one feature, so it solves 5 sets of covariance features, and XGB-RFECV has the smallest RMSE value and the best result. However, the ImprovedRFECV algorithm solves four groups of covariance features, which makes the prediction accuracy of all four models increase. From Figure 5(d), it can be seen that the RMSE values of LGBM, GBDT, and RF predicted by using the best subset of features obtained by the ImprovedRFECV algorithm are all smaller than the RMSE values of prediction by using the best feature best obtained by the RFECV algorithm. Among them, it can be seen from Table 2 that the prediction accuracy of LGBM-ImprovedRFECV is 22% higher than that of LGBM-RFECV. There is a set of covariance features in the power load dataset with 4 features. The

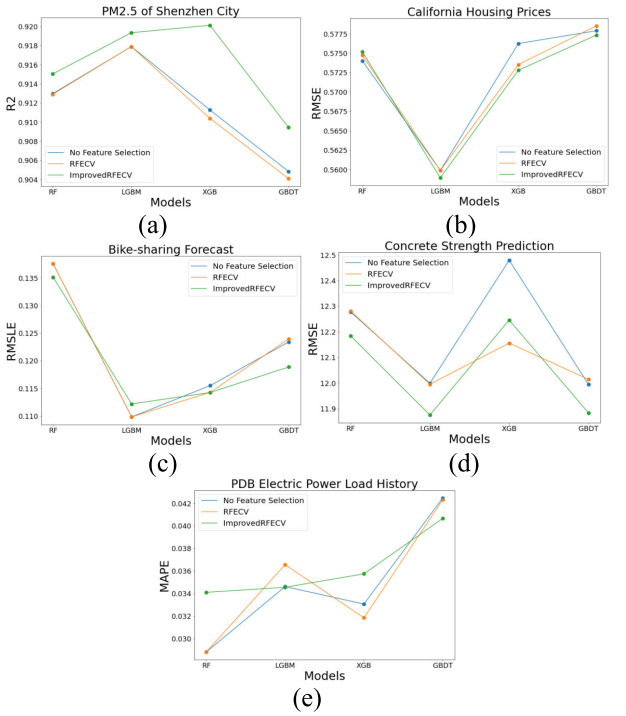

FIGURE 5. Experimental results on the predictive performance of various models combined with diverse feature selection algorithms across multiple domains.

best feature subsets of LGBM-RFECV, XGB-RFECV, and RF-RFECV all contain these 4 features, and the best feature subset of GBDT-RFECV contains 3 features in this group, while the ImprovedRFECV algorithm eliminates 2 features in this group, which more effectively mitigates the covariance problem.

Figure 5(a) shows that all four models achieve the highest R2 when using the best subset of features selected by the ImprovedRFECV algorithm for prediction. Particularly, the XGB-ImprovedRFECV model demonstrates a higher effectiveness. Figure 5(b) demonstrates that combining LGBM, XGB, and GBDT with the best subset of features selected by the ImprovedRFECV algorithm leads to the lowest RMSE values. In contrast, RF achieves its smallest RMSE value without feature selection.

The L1, and L2 regularization introduced in the ImprovedRFECV algorithm can effectively mitigate the covariance between features. And ImprovedRFECV algorithm also introduces stability evaluation, which enables a more comprehensive evaluation of the quality of feature subsets. The integrated learning framework is also used to improve the generalization ability of the algorithm. Therefore, the best feature subset obtained through the ImprovedRFECV algorithm can result in superior predictive accuracy when applied to model predictions.

VI. CONCLUSION

This paper proposes the ImprovedRFECV algorithm, a novel wrapped feature selection approach that enhances the accuracy and generalization capability of machine learning

models. The algorithm engages in the following steps: 1) First, a randomly selected portion of the original data is used as a training and testing set to train the models and evaluate their performance. 2) The above process is repeated several times, and the variance of scores obtained from different models for each feature is summed up to compute the stability score for each feature. 3) To mitigate overfitting, the importance of each feature is assessed comprehensively by using L1 and L2 regularization term scores with different regression models. 4) A both-end expansion removal strategy employed to mitigate the strong covariance between features, while maintaining flexibility in the number of selected.

The algorithm underwent testing on five datasets encompassing various domains, exhibiting superior performance in comparison to the RFECV algorithm. In particular, the algorithm achieved a 13-percentage point improvement in the highest performance, and the average size of the optimal feature subset was reduced by two features. These emphasize the robustness and generalization capability of the ImprovedRFECV algorithm, highlighting its potential for application in various domains with diverse datasets.

REFERENCES

[1] H. H. Inbarani, M. Bagyamathi, and A. T. Azar, "A novel hybrid feature selection method based on rough set and improved harmony search," *Neural Comput. Appl.*, vol. 26, no. 8, pp. 1859–1880, Nov. 2015.

[2] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, Univ. Waikato, Hamilton, New Zealand, 1999.

[3] T. Liu, R. Hu, and Y. Zhu, "Completed sample correlations and feature dependency-based unsupervised feature selection," *Multimedia Tools Appl.*, vol. 82, no. 10, pp. 15305–15326, Apr. 2023.

[4] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, May 2003.

[5] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Comparative study of attribute selection using gain ratio and correlation based feature selection," *Int. J. Inf. Technol. Knowl. Manag.*, vol. 2, no. 2, pp. 271–277, 2010.

[6] H.-H. Hsu, C.-W. Hsieh, and M.-D. Lu, "Hybrid feature selection by combining filters and wrappers," *Exp. Syst. Appl.*, vol. 38, no. 7, pp. 8144–8150, Jul. 2011.

[7] Y. Chen, "Lightweight intrusion detection system based on feature selection," *J. Softw.*, vol. 18, no. 7, p. 1639, 2007.

[8] E. Hancer, B. Xue, and M. Zhang, "Differential evolution for filter feature selection based on information theory and feature ranking," *Knowl.-Based Syst.*, vol. 140, pp. 103–119, Jan. 2018.

[9] R. Ruiz, J. C. Riquelme, and J. S. Aguilar-Ruiz, "Incremental wrapper-based gene selection from microarray data for cancer classification," *Pattern Recognit.*, vol. 39, no. 12, pp. 2383–2392, Dec. 2006.

[10] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, pp. 273–324, Dec. 1997.

[11] E. Ziv, O. Tymofiyeva, D. M. Ferriero, A. J. Barkovich, C. P. Hess, and D. Xu, "A machine learning approach to automated structural network analysis: Application to neonatal encephalopathy," *PLoS ONE*, vol. 8, no. 11, Nov. 2013, Art. no. e78824.

[12] M. Johannes, J. C. Brase, H. Fröhlich, S. Gade, M. Gehrmann, M. Fälth, H. Sultmann, and T. Beißbarth, "Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients," *Bioinformatics*, vol. 26, no. 17, pp. 2136–2144, Sep. 2010.

[13] T. Almutiri and F. Saeed, "A hybrid feature selection method combining Gini index and support vector machine with recursive feature elimination for gene expression classification," *Int. J. Data Mining, Model. Manag.*, vol. 14, no. 1, pp. 41–62, 2022.

[14] M. Lee and J.-H. Lee, "A robust fusion algorithm of LBP and IMF with recursive feature elimination-based ECG processing for QRS and arrhythmia detection," *Int. J. Speech Technol.*, vol. 52, no. 1, pp. 939–953, Jan. 2022.

[15] E. Gysels, P. Renevey, and P. Celka, "SVM-based recursive feature elimination to compare phase synchronization computed from broadband and narrowband EEG signals in brain–computer interfaces," *Signal Process.*, vol. 85, no. 11, pp. 2178–2189, Nov. 2005.

[16] Y. Chen, J. Bi, and J. Z. Wang, "MILES: Multiple-instance learning via embedded instance selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1931–1947, Dec. 2006.

[17] K. Y. Aram, S. S. Lam, and M. T. Khasawneh, "Linear cost-sensitive max-margin embedded feature selection for SVM," *Exp. Syst. Appl.*, vol. 197, Jul. 2022, Art. no. 116683.

[18] S. Zhang, D. Cheng, R. Hu, and Z. Deng, "Supervised feature selection algorithm via discriminative ridge regression," *World Wide Web*, vol. 21, no. 6, pp. 1545–1562, Nov. 2018.

[19] V. Bruni, M. L. Cardinali, and D. Vitulano, "A short review on minimum description length: An application to dimension reduction in PCA," *Entropy*, vol. 24, no. 2, p. 269, Feb. 2022.

[20] H. Liu, M. Zhou, and Q. Liu, "An embedded feature selection method for imbalanced data classification," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 3, pp. 703–715, May 2019.

[21] R. Hu, D. Cheng, W. He, G. Wen, Y. Zhu, J. Zhang, and S. Zhang, "Low-rank feature selection for multi-view regression," *Multimedia Tools Appl.*, vol. 76, no. 16, pp. 17479–17495, Aug. 2017.

[22] S. Roweis, "EM algorithms for PCA and SPCA," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 10, 1997, pp. 626–632.

[23] B. A. Draper, K. Baek, M. S. Bartlett, and J. R. Beveridge, "Recognizing faces with PCA and ICA," *Comput. Vis. Image Understand.*, vol. 91, nos. 1–2, pp. 115–137, Jul. 2003.

[24] V. D. Calhoun, J. Liu, and T. Adali, "A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data," *NeuroImage*, vol. 45, no. 1, pp. 163–172, Mar. 2009.

[25] R. Setiono and H. Liu, "Neural-network feature selector," *IEEE Trans. Neural Netw.*, vol. 8, no. 3, pp. 654–662, May 1997.

[26] R. Hu, L. Zhang, and J. Wei, "Adaptive Laplacian support vector machine for semi-supervised learning," *Comput. J.*, vol. 64, no. 7, pp. 1005–1015, Aug. 2021.

[27] Y. Wang, R. Niu, G. Lin, Y. Xiao, H. Ma, and L. Zhao, "Estimate of soil heavy metal in a mining region using PCC-SVM-RFECV-AdaBoost combined with reflectance spectroscopy," *Environ. Geochemistry Health*, vol. 45, no. 12, pp. 9103–9121, Dec. 2023.

[28] H. Liang, J. Wu, H. Zhang, and J. Yang, "Two-stage short-term power load forecasting based on RFECV feature selection algorithm and a TCN–ECA–LSTM neural network," *Energies*, vol. 16, no. 4, p. 1925, Feb. 2023.

[29] G. Veljanovski, P. Popovski, M. Atanasovski, and M. Kostov, "Implementation of neural networks and feature selection for short term load forecast," in *Proc. 57th Int. Scientific Conf. Inf., Commun. Energy Syst. Technol. (ICEST)*, Jun. 2022, pp. 1–4.

[30] J. Sung, S. Han, H. Park, S. Hwang, S. J. Lee, J. W. Park, and I. Youn, "Classification of stroke severity using clinically relevant symmetric gait features based on recursive feature elimination with cross-validation," *IEEE Access*, vol. 10, pp. 119437–119447, 2022.

[31] C. I. Ossai, D. Rankin, and N. Wickramasinghe, "Preadmission assessment of extended length of hospital stay with RFECV-ETC and hospital-specific data," *Eur. J. Med. Res.*, vol. 27, no. 1, pp. 1–16, Dec. 2022.

[32] M. Amakrane, G. Khaissidi, M. Mrabti, A. Ammour, B. Faouzi, and G. Aboulem, "Feature selection of Arabic online handwriting using recursive feature elimination for Parkinson's disease diagnosis," in *Proc. ES Web Conf.*, vol. 351, 2022, Art. no. 01044.

[33] X. Hou, Z. Quan, A. Aierken, D. Zhao, S. Ji, J. Ni, K. Liu, and H. Qing, "Machine-learning based strategy identifies a robust protein biomarker panel for Alzheimer's disease in cerebrospinal fluid," 2023.

[34] T. A. Assegie, P. K. Rangarajan, N. K. Kumar, and D. Vigneswari, "An empirical study on machine learning algorithms for heart disease prediction," *IAES Int. J. Artif. Intell. (IJ-AI)*, vol. 11, no. 3, p. 1066, Sep. 2022.

[35] J. Tong, C. Song, T. Tong, X. Zong, Z. Liu, S. Wang, L. Tan, Y. Li, and Z. Chang, "Design and optimization of electronic nose sensor array for real-time and rapid detection of vehicle exhaust pollutants," *Chemosensors*, vol. 10, no. 12, p. 496, Nov. 2022.

[36] N. Bamunuarachchi and C. D. Silva, "Developing of NPA predictive model for pawning advances in sri Lankan banking industry," in *Proc. 2nd Int. Conf. Innov. Technol. (INOCON)*, Mar. 2023, pp. 1–7.

[37] Y. W. Chang and C. J. Lin, "Feature ranking using linear SVM," in *Proc. PMLR*, 2008, pp. 53–64.

[38] A. Z. Mustaqim, S. Adi, Y. Pristyanto, and Y. Astuti, "The effect of recursive feature elimination with cross-validation (RFECV) feature selection algorithm toward classifier performance on credit card fraud detection," in *Proc. Int. Conf. Artif. Intell. Comput. Sci. Technol. (ICAICST)*, Jun. 2021, pp. 270–275.
[39] H. Fanaee-T and J. Gama, "Event labeling combining ensemble detectors and background knowledge," *Prog. Artif. Intell.*, vol. 2, nos. 2–3, pp. 113–127, Jun. 2014.
[40] W. Reade and A. Chow. (2023). *Regression With a Tabular California Housing Dataset*. [Online]. Available: https://kaggle.com/competitions/playground-series-s3e1
[41] W. Reade and A. Chow. (2023). *Regression With a Tabular Concrete Strength Dataset*. [Online]. Available: https://kaggle.com/competitions/playground-series-s3e9
[42] H. Karimian, Q. Li, C. Wu, Y. Qi, Y. Mo, G. Chen, X. Zhang, and S. Sachdeva, "Evaluation of different machine learning approaches to forecasting PM2.5 mass concentrations," *Aerosol Air Quality Res.*, vol. 19, no. 6, pp. 1400–1410, 2019.
[43] T. Tu, Y. Su, Y. Tang, G. Guo, W. Tan, and S. Ren, "SHFW: Second-order hybrid fusion weight-median algorithm based on machining learning for advanced IoT data analytics," *Wireless Netw.*, pp. 1–13, Jun. 2023.
[44] J. Manasa, R. Gupta, and N. S. Narahari, "Machine learning based predicting house prices using regression techniques," in *Proc. 2nd Int. Conf. Innov. Mech. Ind. Appl. (ICIMIA)*, Mar. 2020, pp. 624–630.
[45] H. K. Alfares and M. Nazeeruddin, "Electric load forecasting: Literature survey and classification of methods," *Int. J. Syst. Sci.*, vol. 33, no. 1, pp. 23–34, Jan. 2002.
[46] I.-C. Yeh, "Analysis of strength of concrete using design of experiments and neural networks," *J. Mater. Civil Eng.*, vol. 18, no. 4, pp. 597–604, Aug. 2006.
[47] H.-G. Ni and J.-Z. Wang, "Prediction of compressive strength of concrete by neural networks," *Cement Concrete Res.*, vol. 30, no. 8, pp. 1245–1250, Aug. 2000.

TIANYI TU received the master's degree in computer application technology from Xiangtan University, China, in 2005. He is currently a Lecturer with the Hunan University of Arts and Sciences. His research interests include machine learning, software engineering, and algorithm application.

YE SU is currently pursuing the bachelor's degree with the College of Computer and Electrical Engineering, Hunan University of Arts and Science. His research interests include feature selection algorithms, hybrid fusion algorithms, and automated feature engineering algorithms. He became a member of CCF, in 2022, and IEEE CS, in 2023.

YAYUAN TANG received the Ph.D. degree in computer science and technology from Central South University. She is currently an Associate Professor with the School of Information Engineering, Hunan University of Science and Engineering. Her research interests include intelligent computing, resource scheduling, and big data retrieval.

WENXUE TAN was born in 1973. He received the Ph.D. degree from the College of Computer Science, Beijing University of Technology, in 2016. He is currently a Professor and a Senior Engineer with the Hunan University of Arts and Science. He has published over 38 research articles, 19 of which were indexed by EI Compendex or SCI database, and eight of which were referred by the Chinese Science Citation Database. His current research interests include agriculture information technology, artificial intelligence, and cloud information security.

SHENG REN received the Ph.D. degree in computer science and engineering from Central South University, China, in 2022. He is currently an Associate Professor with the Hunan University of Arts and Sciences. His research interests include big data, image and video super-resolution, video analysis, and understanding.