Machine Learning with Applications 6 (2021) 100170

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Assessing feature selection method performance with class imbalance data

Surani Matharaarachchi [a,*], Mike Domaratzki [b,1], Saman Muthukumarana [a]

[a] Department of Statistics, University of Manitoba, Winnipeg, MB, R3T 2N2, Canada
[b] Department of Computer Science, University of Manitoba, Winnipeg, MB, R3T 2N2, Canada

ARTICLE INFO

Keywords:
Feature selection
Informative feature
Recursive feature elimination
Principal component loading

ABSTRACT

Identifying the most informative features is a crucial step in feature selection. This paper focuses primarily on wrapper feature selection methods designed to detect important features with F1-score as the target metric. As an initial step, most wrapper methods order features according to importance. However, in most cases, the importance is defined according to the classification method used and varies with the characteristics of the data set. Using synthetically simulated data, we examine four existing feature ordering techniques to find the most desirable and the most effective ordering mechanism to identify informative features. Using the results, an improved method is suggested to extract the most informative feature subset from the data set. The method uses the sum of absolute values of the first k principal component loadings to order the features where k is a user-defined application-specific value. It also applies a sequential feature selection method to extract the best subset of features. We further compare the performance of the proposed feature selection method with results from the existing Recursive Feature Elimination (RFE) by simulating data for several practical scenarios with a different number of informative features and different imbalance rates. We also validate the method using a real-world application on several classification methods. The results based on the accuracy measures indicate that the proposed approach performs better than the existing feature selection methods.

1. Introduction

Feature selection determines the features which should be included in a model. With the curse of dimensionality (Bellman, 1957), data are becoming increasingly high-dimensional, and feature selection is becoming one of the most critical topics to consider. A perfect feature selection method should choose the most informative features and eliminate the less informative features. Therefore, it should primarily be focused on removing non-informative features from the model (Kuhn, 2013) and achieving higher accuracy with the most informative features. The challenge is that it is computationally demanding, time-consuming, and not practical to compare all the combinations of features to determine which combination achieves the highest accuracy. Therefore, feature selection techniques need to address these concerns while achieving some significant advantages.

When we select fewer features before applying them to the predictive classification models, it decreases computational time and improves model interpretability (Miche et al., 2007). Statistically, it is more convenient and attractive to estimate fewer parameters, and it also reduces the negative impact of non-informative features.

Mainly, three categories of feature selection methods are introduced in the literature: filter, wrapper, and embedded methods (Lal et al., 2006; Stańczyk, 2015). Filter methods measure the relevance of features by their correlation with the dependent variable; hence, only features with meaningful relationships would be included in a classification model. On the other hand, wrapper methods measure the usefulness of a subset of features by actually training a model on it (Saeys et al., 2007). They evaluate multiple subsets, adding and/or removing features to find the optimal combination that maximizes overall model performance. Some wrapper methods perform this evaluation with different randomly selected subsets, using a cross-validation (CV) method. Forward Feature Selection, Backward Feature Elimination (Weisberg, 2005), and Recursive Feature Elimination (RFE) (Guyon et al., 2002) are typical examples of commonly used wrapper methods. The third category, embedded methods, is quite similar to wrapper methods. Although the embedded methods also optimize the objective

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author.
E-mail addresses: matharas@myumanitoba.ca (S. Matharaarachchi), mdomarat@uwo.ca (M. Domaratzki), saman.muthukumarana@umanitoba.ca (S. Muthukumarana).
[1] Present Address: Department of Computer Science, Western University, London, ON, N6A 5B7, Canada.

https://doi.org/10.1016/j.mlwa.2021.100170
Received 28 April 2021; Received in revised form 20 September 2021; Accepted 21 September 2021
Available online 5 October 2021

function or performance of a learning algorithm or model, they also uses an intrinsic model building metric during learning. L1 (LASSO) regularization (Tibshirani, 1996) and decision tree algorithm (Breiman et al., 1984) are commonly known embedded methods.

For most wrapper methods, the features are ordered according to their importance as an initial step. Although wrapper methods are computationally more intensive than filter methods, they have significant advantages over filter methods such as higher classification accuracy, interaction with classifiers, and model feature dependencies (Kumari & Swarnkar, 2011). This paper focuses on examining the impact of the number of informative features and imbalance rates on feature ordering and feature selection techniques. Ultimately, it proposes a modified wrapper feature selection method, which also examines subsets of features to improve accuracy on fewer features.

Although there are several feature selection techniques in the literature, they behave uniquely with varying data sets. In particular, with standard wrapper feature selection techniques, different classification models select features differently, even for the same data set. Most feature selection techniques may not accurately determine all informative features when the class sizes are drastically different. Hence, the resulting output may not be the one anticipated. Therefore, identifying the most suitable feature ordering technique and most informative feature subset with different class imbalance levels are significant concerns requiring a solidified solution.

Principal component analysis (PCA) is a statistical technique for reducing the dimensionality of high-dimensional data sets. For a data set $\mathbf{X}$ of dimension n by p, PCA attempts to find linear orthogonal combinations of the columns of $\mathbf{X}$ with maximum variance such that $\sum_i A_i X_i = \mathbf{XA}$. After implementing PCA on the data set, the original features will turn into principal components, linear combinations of the original features. Hence, principal components are not as readable and interpretable as original features. But, in this paper, we consider the PC loadings, the weights of the features of each linear combination, to avoid this interpretability issue.

To verify a new general approach for choosing the most important features in any situation, we train our binary prediction models using five commonly-used classification techniques, i.e., Logistic Regression (LOGIT) (Hastie et al., 2009; McCullagh & Nelder, 1989), Linear Support Vector Machine with linear kernel (SVM_lin) (Cortes & Vapnik, 1995; Xia & Jin, 2008), Decision Tree (DT) (Breiman et al., 1984; Guo et al., 2002), Random Forest (RFC) (Breiman, 2001), and Light Gradient Boosting (LGBM) (Friedman, 2001). .

These classification models are then evaluated using commonly-used performance measures (e.g., F1-score). Recursive feature elimination (RFE) technique is used to compare the proposed method with, and the synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002) is used as the class re-balancing technique for the real-world application data. To cover practical scenarios, we synthetically generated data using the "make classification" library in python scikit-learn.datasets (Pedregosa et al., 2011). However, to better understand the impact of class re-balancing and feature selection on binary classification models, we attempt to address two research questions: finding the best feature ordering technique and determining which method extracts the best informative feature subset.

The rest of this paper is organized as follows. Section 2 will detail the data preparation, methods used in the study, and experimental design. Section 3 presents the results of our simulation studies. We illustrate the results in real-world applications and interpret the results in Section 4. Section 5 of this paper is evaluated with a discussion of its contributions and limitations and future research directions, whereas 6 resolves with a conclusion.

2. Methods and experimental design

We began with a simulation study to examine the effect of informative and non-informative features on feature ordering techniques. We simulated data and fixed the total number of features to be 30. In the data simulation, each class is formed of several Gaussian clusters, each located around the vertices of a hypercube in a subspace of dimension equal to the number of informative features. In this study, the number of classes was two, and there was only one cluster per class. Informative features are drawn independently from Normal(0, 1) distribution for each cluster and then combined as random linear combinations within each cluster to add covariance (Pedregosa et al., 2011). The remaining non-informative features are filled with random noise. Data sets were generated by increasing the number of informative features from 1 to the total number of features. Since the number of features is fixed at 30, the remaining features are non-informative.

The sample size and the imbalance rate of the data set were changed according to the problem definition. For model validation purposes, each data set was divided into two parts, training (75%) and testing (25%). Finally, we obtained the accuracy measures for classification methods combined with the imbalance rate, sample size, and the number of informative features given in the data set.

2.1. What is the best feature ordering technique?

By ordering features, we mean placing them in an order according to their importance to identify the most informative features. We use four different feature selection methods to compare the feature ordering behavior.

1. **Summation of the absolute values of principal component loadings**
   After applying principal component analysis (PCA) to the data set, we are interested in understanding the relationship of the original variables to the principal components using PC loadings (Dunteman, 1989). PC loadings are the coefficients of the linear combination of the original variables from which the principal components (PCs) are constructed. In PCA, given a mean-centered data set $\mathbf{x}$ with n sample and p variables, the first k principal components, PC_1, PC_2, …, PC_k respectively are given by the linear combination of the original variables $\underline{X}_1, \underline{X}_2, \cdots, \underline{X}_p$,
   $$PC_1 = w_{11}\underline{X}_1 + w_{12}\underline{X}_2 + \ldots + w_{1p}\underline{X}_p$$
   $$PC_2 = w_{21}\underline{X}_1 + w_{22}\underline{X}_2 + \ldots + w_{2p}\underline{X}_p$$
   $$\vdots$$
   $$PC_k = w_{k1}\underline{X}_1 + w_{k2}\underline{X}_2 + \ldots + w_{kp}\underline{X}_p.$$
   We then compute the sum of the absolute values of the two PC loadings for each feature and order features accordingly. That is for $\underline{X}_i$, it is $\sum_{j=1}^{k} |w_{ji}|$, where $i = 1, \ldots, p$.
2. **Univariate feature selection (ANOVA F value classification)**
   Analysis of variance (ANOVA) is used to analyze the differences among group means in a sample. The F-test is a statistical test used to compare the factors by decomposing the total variation. For example, in one-way or single-factor ANOVA, statistical significance is tested by comparing the F-test statistic,
   $$F = \frac{\text{variability between groups}}{\text{variability within groups}}$$
   These test results can be used in feature selection by removing features independent of the target variable (Kuhn, 2013). In our analysis, we order features according to F values (p values) to identify the most informative features.
3. **Absolute correlation of features with the response variable**
   We consider the point biserial correlation to measure the relationship between a binary variable, Y, and a continuous variable, X. This coefficient also varies between −1 and +1 where 0 implies no correlation. The absolute value of a point biserial

correlation coefficient $|r_{bp}|$ describes the magnitude of the relationship between two variables and uses a t-test with $n-1$ degree of freedom. If we divide the data set into two groups, according to 0 and 1 in Y, the absolute value of the point biserial correlation can be expressed in the form,

$$|r_{bp}| = \left| \frac{M_1 - M_0}{S_{n-1}} \sqrt{\frac{n_0 n_1}{n(n-1)}} \right|$$

where,

$$S_{n-1} = \sqrt{\frac{1}{n-1} \sum_{r=1}^{n} (X_i - \bar{X})^2}$$

Here, M_0 and M_1 are means on the continuous variable X for all data points in group 0 and 1 respectively, S_{n-1} is the standard deviation of the metric observations, n_0 and n_1 are the number of observations for each group and n is the total number of observations (Lev, 1949; Tate, 1954).

4. **Classification model-based feature importance**
We now consider the feature importance, which was directly obtained from the classification model trained. More specifically, we look at two main types of more advanced feature importance; they are:

   (a) Feature importance from model coefficients (Tsuruoka et al., 2009): Linear machine learning algorithms fit a model where the prediction is the weighted sum of the original features and these weights (Coefficients) can be used directly to measure the feature importance. Examples include logistic regression and support vector machine with linear kernel.
   (b) Feature importance from decision trees (Breiman et al., 1984): Decision tree algorithms like classification and regression trees (CART) offer importance scores based on minimizing the criterion used to select split points, like Gini or entropy. Examples of such models include decision trees, random forest, and gradient boosting algorithms.

2.2. Which method extracts the best informative feature subset?

After identifying the best feature ordering technique, the next challenge is to obtain the most informative feature subset. To achieve this objective, we suggest a better feature selection technique and compare the results with an existing feature selection technique, RFE, which uses model-based feature importance in the initial step. The suggested method uses the sum of absolute values of the principal component loadings and a sequential search method (Peng et al., 2010) to extract the most informative features from the data set. Sequential search usually looks for the optimal feature subset by either adding (or removing) a single feature or a small number of features at a time until the specified criteria are fulfilled (Pudil et al., 1994).

The role of principal component loadings

Principal component analysis (PCA), introduced in 1933 (Hotelling, 1933), has been used in different areas, including the biological, physical, and engineering sciences. The prime purpose of the principal component analysis is to reduce the dimensionality of a multivariate data set and interpret the results by identifying a smaller number of variables. PCA finds the maximum variance in high-dimensional data and projects it onto a smaller dimensional subspace while retaining most of the information. The method requires that correlations be obtained from variables measured on some continuous scale. Also, it assumes a linear relationship between all variables, and large enough sample sizes are required.

However, the most significant disadvantage of PCA is that our original features will be transformed into principal components after implementing PCA on the data set. The principal components are linear combinations of the original features, which are not as interpretable as original features. Hence, we consider using principal component loadings to interpret features and to identify their importance.

Table 1
Model confusion matrix.

		Predictions	
		Class 1	Class 0
Actual	Class 1	TP_{model}	FN_{model}
	Class 0	FP_{model}	TN_{model}

Table 2
Feature selection confusion matrix.

	Selected	Not selected
Informative	TP_{fs}	FN_{fs}
Non-informative	FP_{fs}	TN_{fs}

Suggested method: Principal component loading feature selection method (PCLFS)

The suggested method employs the first k principal components where k is a user-defined application-specific value. We consider only two principal components for the simulation study, assuming the highest sample variances are accumulated in the first few principal components.

The first step is ordering features using the sum of the first two principal component loadings' absolute values. The second is selecting the optimal feature subset, which obtains the maximum F1-score. Starting from the most informative feature, we add features one by one according to the pre-defined order until all features are added. Hence the total number of subsets will equal the number of features in the data set. We validate the models using the testing set at each step (i.e., F1-score) and, in the end, obtain the feature subset which gives the maximum F1-score. Fig. 1 shows the process of the suggested method. Further, to validate the proposed method and justify the behavior with different samples, we used 5-fold cross-validation, and the results will be discussed in Section 3.

Introducing feature selection confusion matrix

Generally, a confusion matrix is a table that can be used to describe the performance of a classification model. Here, since we already know the number of informative and non-informative features in the simulated data set, we introduce a new confusion matrix to check the feature selection performance of each classification method with any feature selection method.

The new feature selection confusion matrix can be defined as in Table 2 where Table 1 is the regular model confusion matrix. In the Tables, the outcomes are: TP = True Positive, FP = False Positive, TN = True Negative, and FN = False Negative.

Then, to evaluate the best performing feature selection method, we use the feature selection correct percentage (balanced accuracy) obtained using the newly introduced confusion matrix. The feature selection correct percentage can be calculated as below,

$$TPR_{fs} = \frac{TP_{fs}}{\text{Total Number of informative features}}$$

$$FPR_{fs} = \frac{FP_{fs}}{\text{Total Number of informative features}}$$

$$TNR_{fs} = \frac{TN_{fs}}{\text{Total Number of non informative features}}$$

$$FNR_{fs} = \frac{FN_{fs}}{\text{Total Number of non informative features}}$$

$$Correct\%_{fs} = \frac{TPR_{fs} + TNR_{fs}}{2}$$

Finally, to evaluate the model performance in the real-world data set, we also use model Precision, Recall, and F1-score which can be calculated as

$$\text{Precision} = \frac{TP_{model}}{TP_{model} + FP_{model}}$$

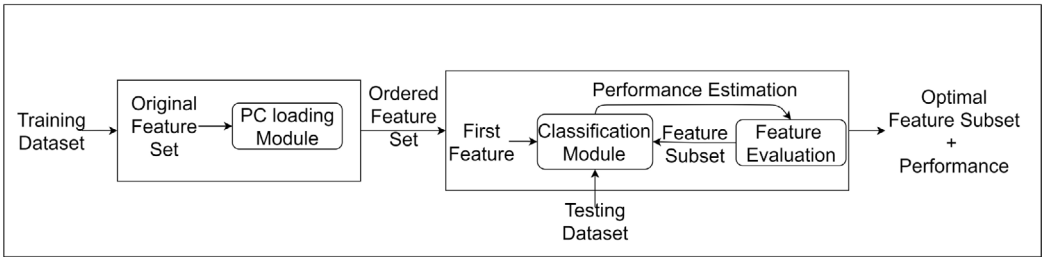

Fig. 1. Principal component loading feature selection method.

$$\text{Recall} = \frac{TP_{model}}{TP_{model} + FN_{model}}$$

$$\text{F1-score} = 2\left(\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}\right)$$

3. Simulation results

To find the best performing method, we simulate one hundred different data sets (with a sample size of 1000 and 30 features) for each imbalance rate. The number of informative features was given from 1 to 30, increasing by 1. For the synthetic comparison, we compare feature ordering only using Logistic regression classification.

3.1. What is the best feature ordering technique?

The objective of this analysis was to identify the best informative feature ordering techniques. As we are using simulated data, we already know the informative features and non-informative features; hence, four feature ordering techniques were applied to determine which would select the informative features first.

Figs. 2, 3, and 4 show three examples of the ordered selection of the features for different imbalance rates from each method when 20 informative features were included in the data set (indicated by dashed lines). The informative features are labeled as 'i_*'. In contrast, the non-informative features indicate 'n_*'. Irrespective of the imbalance rate, the x-axis of the figures clearly shows that the method of having the summation of the absolute values of the first two principal components loadings identified more informative features than the other methods. The ANOVA F classification and the absolute point biserial correlation order features similarly.

To observe the variability of the results, we repeatedly generated 100 data sets for each scenario to meet different practical situations by changing the sample size, the total number of features, number of informative features, and the class imbalanced rate. We applied all four methods for each data set and finally calculated the percentage of selecting informative features using Eq. (1).

$$\text{Percentage of informative features selected} = \frac{S}{I} \tag{1}$$

where S is the average number of informative features selected within the expected range, and I is the number of informative features given. (i.e. expected range is the total number of informative given in the data set).

Figs. 5, 6, and 7 show how the percentage of informative features selected changes with different imbalance rates, number of informative features, and sample sizes. Fig. 8 shows how it changes with a different number of features and informative features when the sample size is 1000 and imbalance rate is 70%:30%. Altogether, it indicates that the performance of the featuring ordering of these methods is robust to the different characteristics of the data set. It is also noted that until four features, the Logit-based method picks the more informative features correctly. When there are more than four informative features in the data set, the sum of the absolute principal loading method picks the most informative features within the expected range compared with the other three methods. It is also shown that the ANOVA F classification and the absolute point biserial correlation behave similarly in all situations.

Since the PC loading method is able to rank the features informatively, there should be a way to extract these informative features in the first phase. Hence, we apply a sequential feature selection technique on the ordered list of features constructed by the PC loading method to obtain the desired feature subset. We choose the Logit absolute coefficient method and recursive feature elimination with the Logit classifier to compare the results of the suggested method. This is described in detail in the next section.

3.2. Which method extracts the best informative feature subset?

To extract the best feature subset, first, we sorted the features using the sum of the absolute values of the first k PC loadings and then obtained F1-scores for each subset of features by fitting a classification model starting from the most important feature and then adding the next important feature until the least important one. Finally, we compared the results with the existing recursive feature elimination technique, which uses classification model-based feature importance to order data.

Fig. 9 shows a comparison of cross-validation F1-scores for each subset selected by the Logit-RFE and F1-score of the Logit-PCLFS subsets in data sets with a different number of informative features. Dashed lines indicate the number of informative features included in each data set. More figures related to this comparison are presented in Appendix A. The results are for one set of simulated data with a sample size of 1000, 70%:30% imbalanced rate, and different informative features. The maximum F1-scores are indicated by points on the lines. In most situations, PCLFS yields a higher F1-score than the RFE. Here, the ordering is not done again in PCLFS, only do the sub-set selection. Implying ordering of these features emphasizes that selected features included the informative ones first in each case. It is also notable that even by looking at the PCLFS line, we can identify the informative feature count in the data set. Until it reaches the total number of informative features, the F1-score for the PCLFS method increases rapidly, and the line becomes stable afterward.

Simulation results

Again to capture the variability of the results, a simulation study was done by applying both PCLFS and Logit-RFE methods on training data. Then we evaluated the prediction results on testing data and recorded the model F1-score and the feature selection correct percentage ($\text{Correct}\%_{\text{fs}}$) in both cases. The process repeated 100 times for three different imbalance rates and three different sample sizes.

According to Figs. 10, 11, and 12, we can see that the PCLFS with Logit classifier gives a higher final F1-score in each situation, and it works extremely well for small sample sizes and highly imbalanced data. The figures on the right-hand side imply that the PCLFS method with Logit-classifier obtains a higher feature selection correct percentage in each situation when the number of informative features in the sample is greater than four.

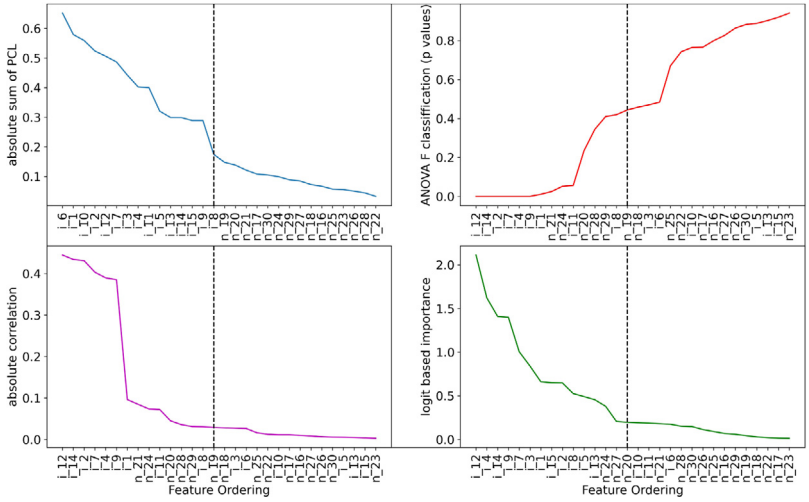

Fig. 2. Example result for comparison of four methods for 50%:50% balanced data: X-axis ordered the features according to the importance given by each method. The black dashed line indicates the number of informative features in the data set.

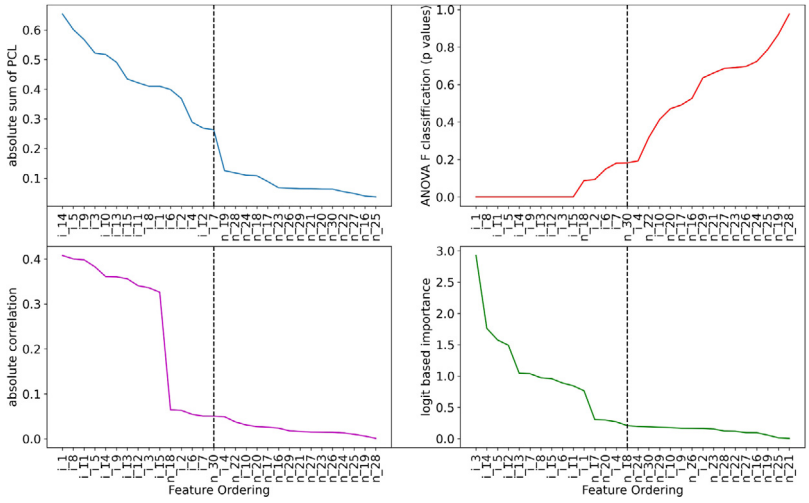

Fig. 3. Example result for comparison of four methods for 70%:30% imbalanced data: X-axis ordered the features according to the importance given by each method. The black dashed line indicates the number of informative features in the data set.

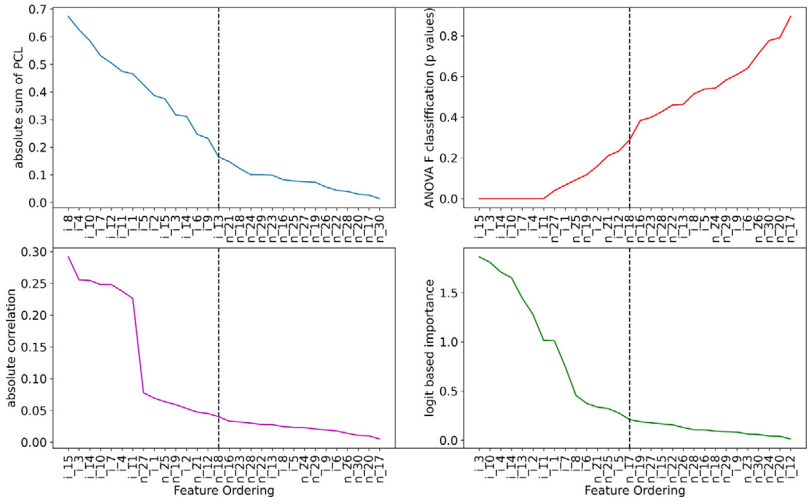

Fig. 4. Example result for comparison of four methods for 90%:10% imbalanced data: X-axis ordered the features according to the importance given by each method. The black dashed line indicates the number of informative features in the data set.

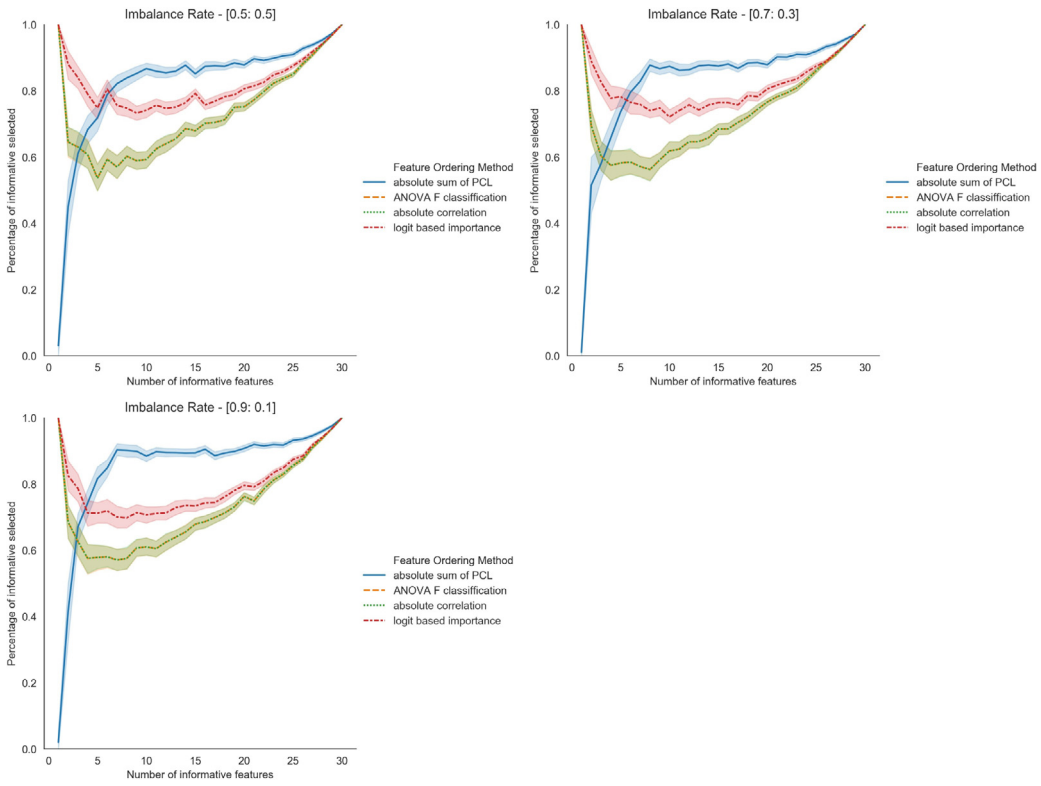

Fig. 5. Mean percentages of informative features selected by each ordering technique in different class imbalance levels with 200 sample sizes. The blue line represents the sum of the absolute values of principal component loadings; the red dashed line indicates Logit model-based feature importance results. The overlapped green and orange dashed lines show the absolute correlation and the ANOVA F value classification results. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

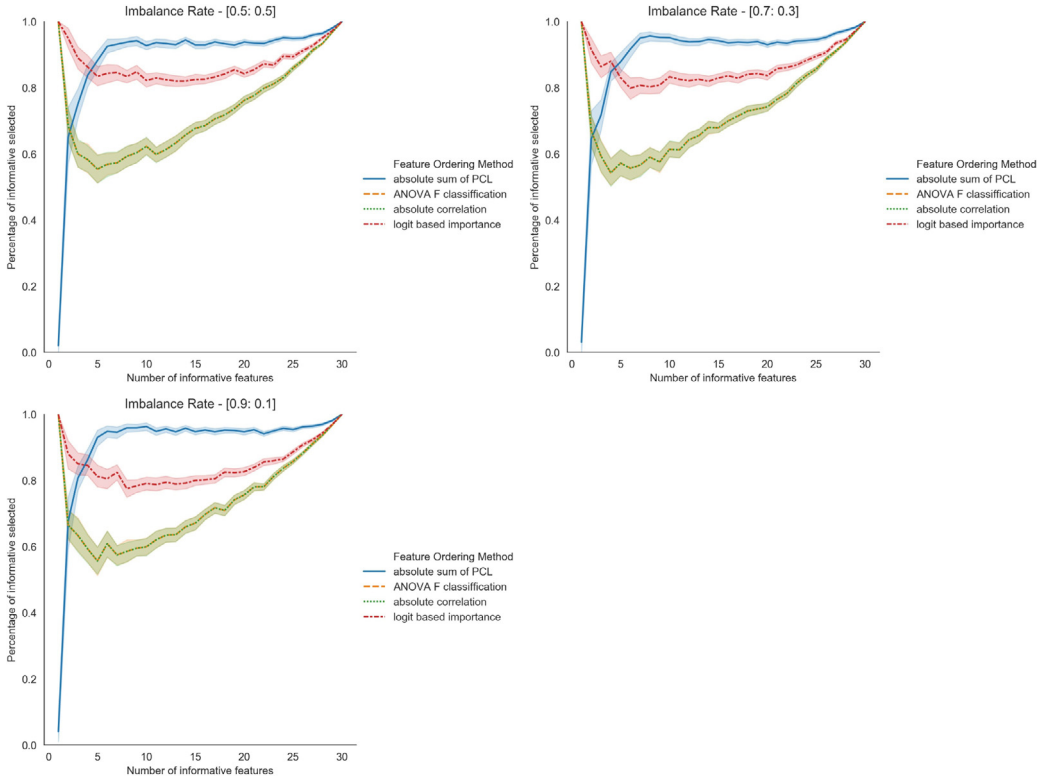

Fig. 6. Mean percentages of informative features selected by each ordering technique in different class imbalance levels with 500 sample sizes. The blue line represents the sum of the absolute values of principal component loadings; the red dashed line indicates Logit model-based feature importance results. The overlapped green and orange dashed lines show the absolute correlation and the ANOVA F value classification results. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

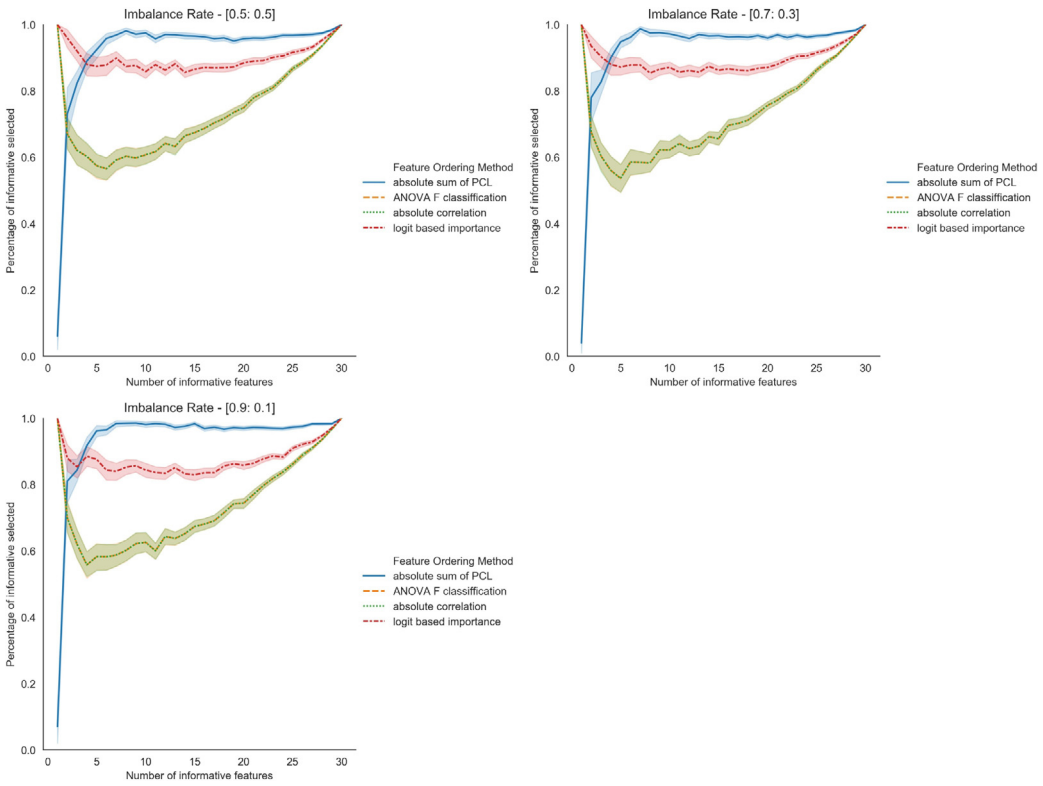

Fig. 7. Mean percentages of informative features selected by each ordering technique in different class imbalance levels with 1000 sample sizes. The blue line represents the sum of the absolute values of principal component loadings; the red dashed line indicates Logit model-based feature importance results. The overlapped green and orange dashed lines show the absolute correlation and the ANOVA F value classification results. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

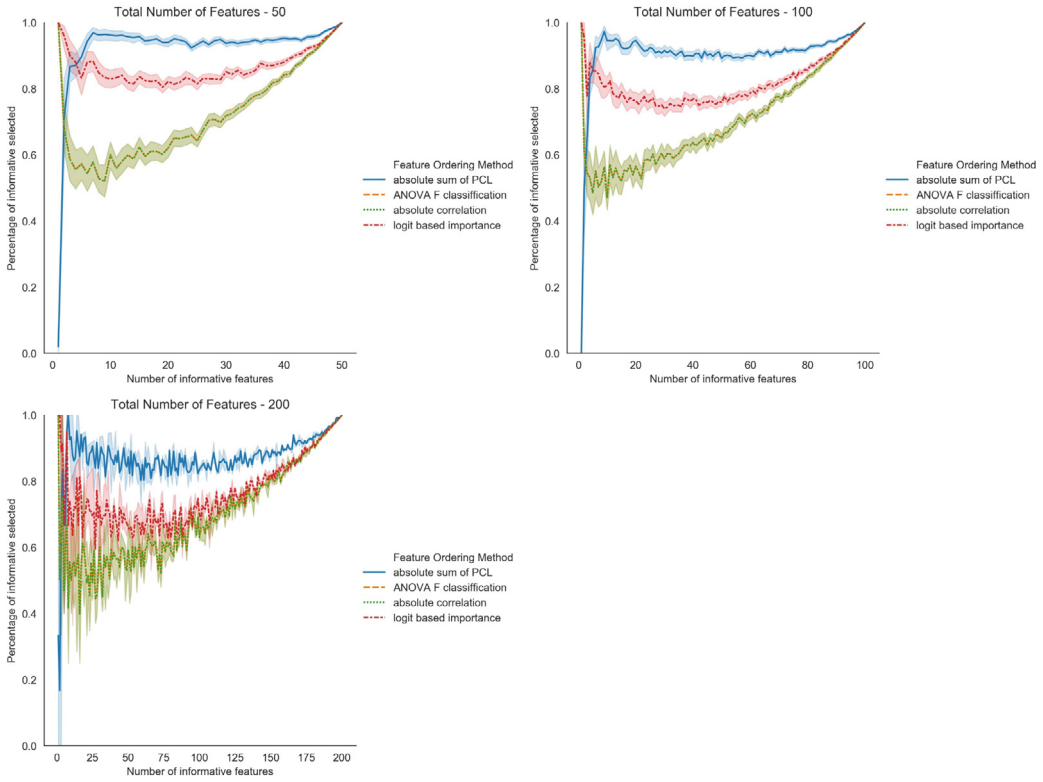

Fig. 8. Mean percentages of informative features selected by each ordering technique in 70%:30% class imbalance level and 1000 sample size with a different number of features. The blue line represents the sum of the absolute values of principal component loadings; the red dashed line indicates Logit model-based feature importance results. The overlapped green and orange dashed lines show the absolute correlation and the ANOVA F value classification results. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

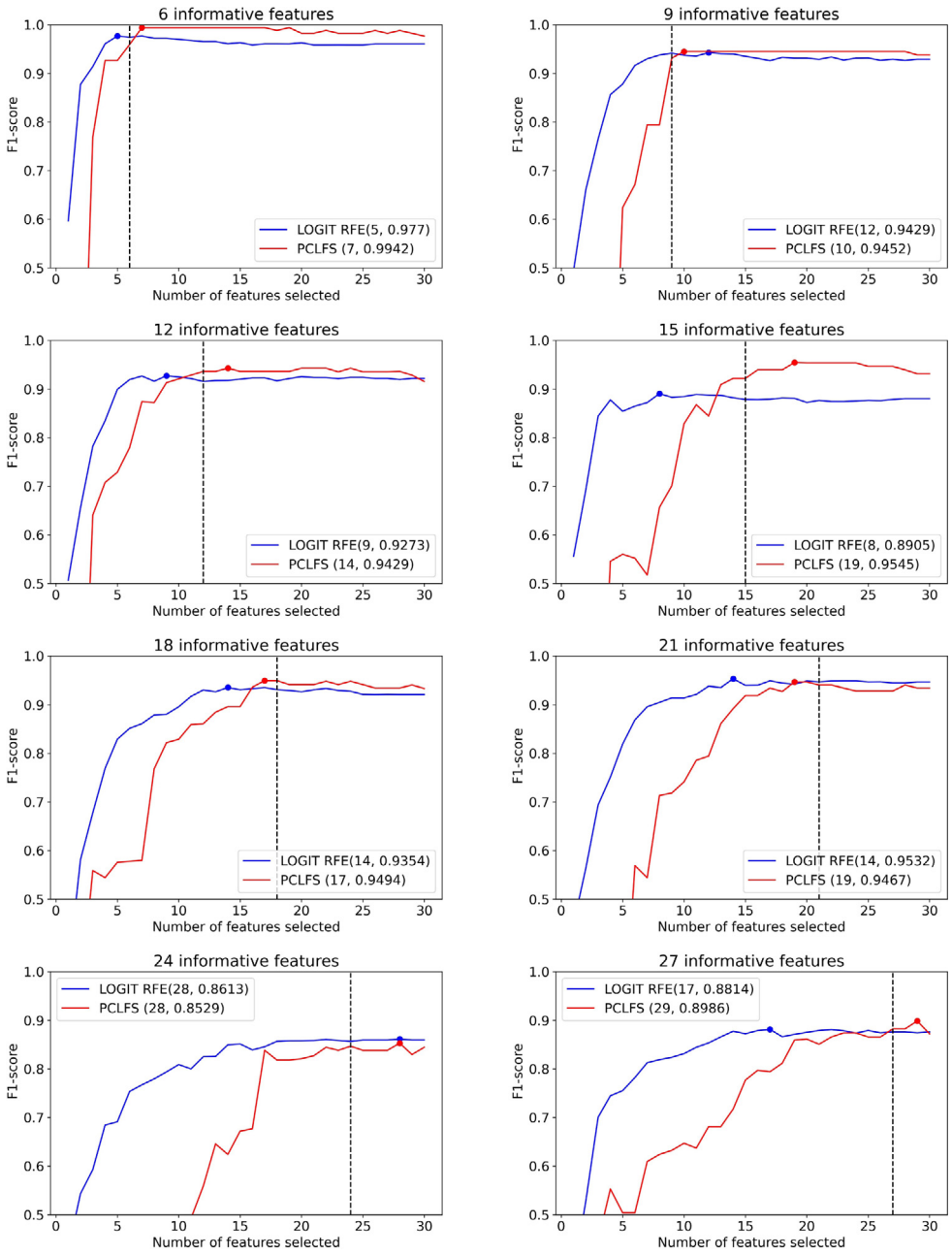

Fig. 9. Comparison between Logit-RFE cross-validation F1-scores and Logit-PCLFS F1-scores for different number of informative features with a sample size of 1000, 70%:30% imbalanced rate. The black dashed line indicates the number of informative features in the data set.

Statistical tests were then conducted on each combination of sample size, imbalance rate, and the number of informative features to assess whether the population medians of the F1-scores given by the two methods differ. Since the distributions and differences are not normally distributed (according to the Shapiro Wilk test), the non-parametric Wilcoxon signed-rank test (Wilcoxon, 1945) was used. It tests the null hypothesis that two related paired samples come from the same distribution with the same medians versus the alternative hypothesis, i.e., two samples come from different distributions where PCLFS has the higher median. According to the p-values of the tests for each scenario, out of 270 tests, 259 tests rejected the null hypothesis. Details of the test which could not reject the null hypotheses are shown in Table 3, and those situations can be identified as the number of informative features in the samples is 1, 2, or 30. Apart from that, we reject the null hypotheses for all the other combinations concluding that the population F1-score median ranks of the PCLFS are greater than the population F1-score median ranks of the Logit-RFE method.

We further conducted a simulation study with cross-validation to perceive the average performance of the proposed models with different training samples. We generated 50 data sets from each situation by changing the class imbalance level and the number of informative features in the data set. The sample size was taken as 1000. The total number of features was 30, where the number of informative features was increased from 1 to 30 in each sample. For each data set, the 5-fold cross-validation F1-scores were calculated, and the averages of the model F1-scores comparison of 5-fold cross-validation runs are shown in Fig. 13. The results reveal that the proposed method works better than different validation sets as well.

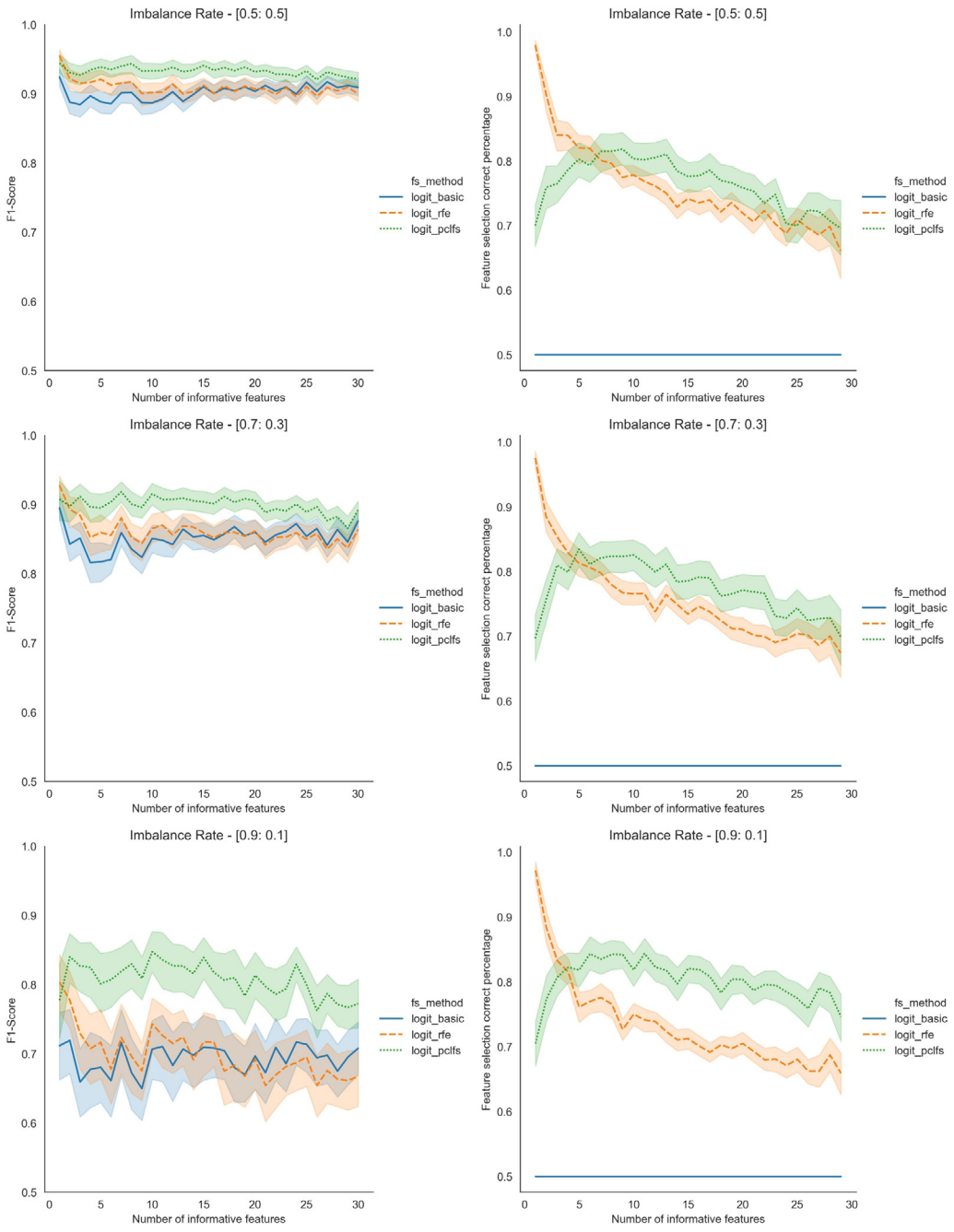

Fig. 10. Final model F1-scores and Feature selection correct percentages for the Logit model when the sample size is 200.

Table 3
Wilcoxon signed-rank test results for not rejecting the null hypothesis.

sample_size	imbalance_rate	n_informative	p_value	Decision
200	0.5: 0.5	1	0.9462	No evidence to reject H0
200	0.7: 0.3	1	0.1934	No evidence to reject H0
200	0.7: 0.3	2	0.0601	No evidence to reject H0
200	0.9: 0.1	1	0.7112	No evidence to reject H0
500	0.5: 0.5	1	0.2590	No evidence to reject H0
500	0.7: 0.3	1	0.7380	No evidence to reject H0
1000	0.5: 0.5	1	0.1026	No evidence to reject H0
1000	0.5: 0.5	30	0.3357	No evidence to reject H0
1000	0.7: 0.3	30	0.3442	No evidence to reject H0
1000	0.9: 0.1	1	0.5322	No evidence to reject H0
1000	0.9: 0.1	30	0.2665	No evidence to reject H0

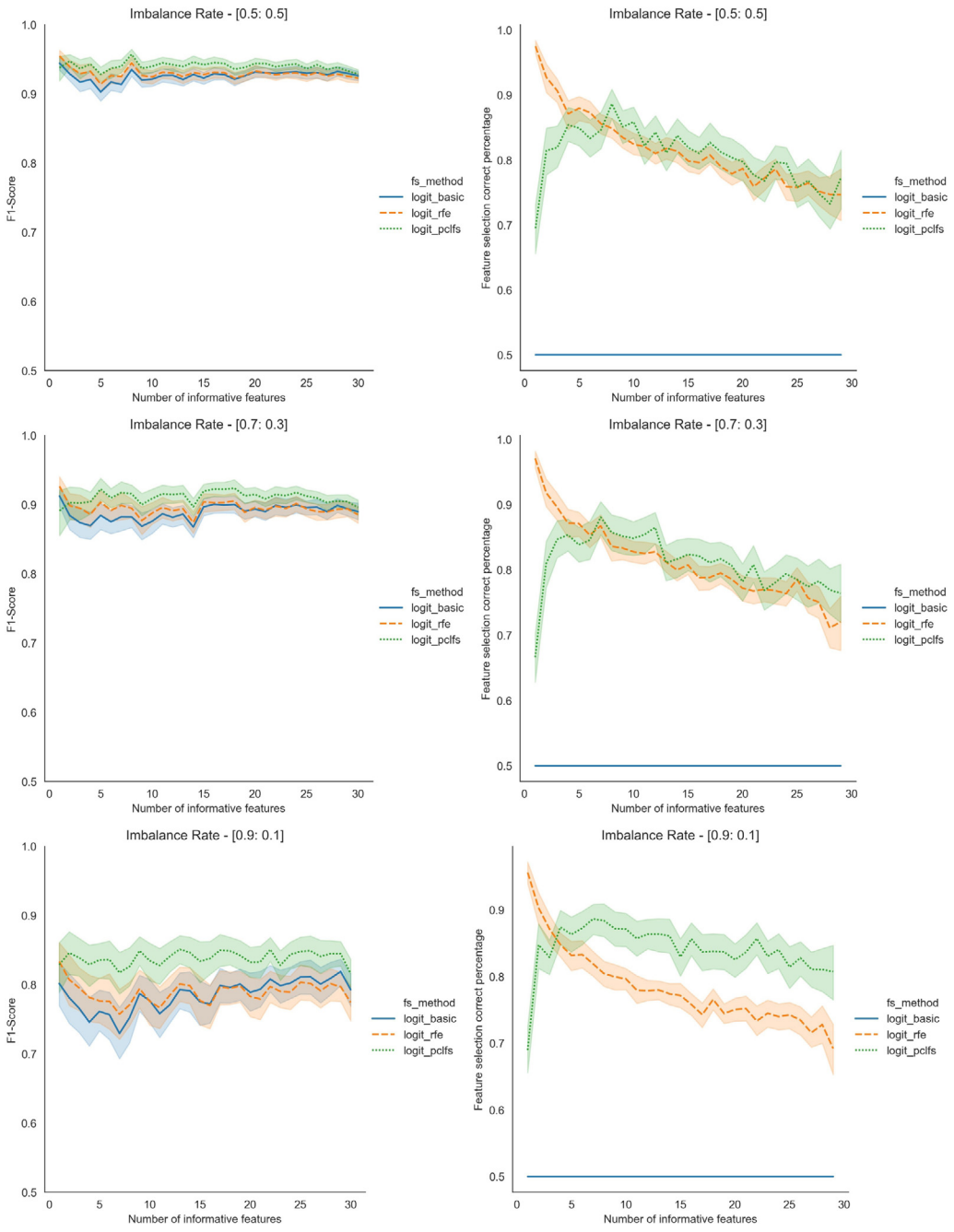

Fig. 11. Final model F1-scores and Feature selection correct percentages for the Logit model when the sample size is 500.

Table 4
Application results for comparing RFE and PCLFS.

SMOTE	Model	RFE				PCLFS			
		#Features	Precision	Recall	F1-score	#Features	Precision	Recall	F1-score
TRUE	Logit	36	0.6154	0.80	0.6957	24	0.6154	0.90	0.6957
	LGBM	27	0.7333	0.55	0.6286	13	0.7857	0.65	0.7027
	Decision Tree	44	0.6250	0.50	0.5556	9	0.7059	0.70	0.6667
	RFC	38	0.6875	0.55	0.6111	42	0.8571	0.70	0.7059
	SVM-Linear	30	0.6522	0.75	0.6977	12	0.7083	0.95	0.7727
FALSE	Logit	30	0.6667	0.40	0.5000	44	0.6923	0.45	0.5455
	LGBM	15	0.6923	0.45	0.5455	15	0.8333	0.50	0.6250
	Decision Tree	27	0.7273	0.40	0.5161	9	1.0000	0.55	0.5946
	RFC	9	0.7143	0.25	0.3704	11	1.0000	0.30	0.4444
	SVM-Linear	21	0.7143	0.50	0.5882	37	0.9000	0.60	0.6316

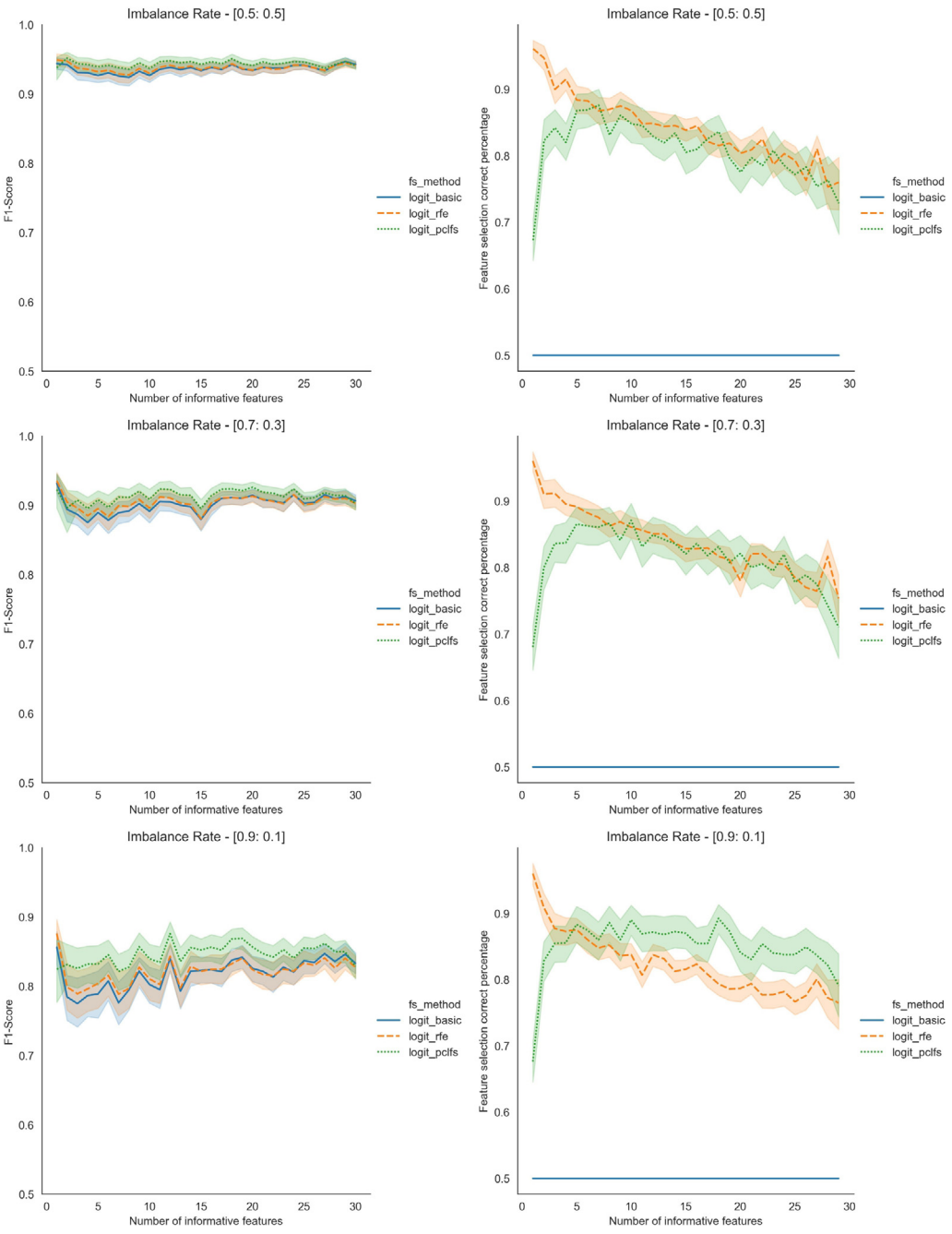

Fig. 12. Final model F1-scores and Feature selection correct percentages for the Logit model when the sample size is 1000.

4. Application

4.1. SPECTF heart data

To analyze the behavior of models on a real-world data set, we consider the publicly available Single-Photon Emission Computed Tomography heart data set (SPECTF) (Krzysztof et al., 1997; Kurgan et al., 2001), which describes diagnosing cardiac abnormalities using SPECT. SPECT is an imaging technique where a radioisotope is used to produce 3D images of a patient using gamma rays (Bruyant, 2002). The data set has classified each patient into two categories: normal and abnormal, by considering the diagnosis of images. This data consists of binary class imbalanced data with a higher number of numerical features and fewer instances.

The data set consists of 267 SPECT image sets (patients), which were processed to extract features that summarize the original SPECT images. As a result, 44 continuous feature patterns were created for each patient. Hence, it has 267 instances that are described by 45 attributes (44 continuous and 1 binary class). We divided the data set into two groups, 75% training samples and 25% test samples. The class-imbalanced rate for the data set is 79.4%:20.6%, where the minority class represents the abnormal patients. The data classification with the first two principal components is shown in Fig. 14.

Then we apply Synthetic Minority Oversampling Technique (SMOTE) to handle imbalance data to achieve higher accuracy in classification models. SMOTE aims to balance class distribution by randomly increasing minority class examples by creating similar instances. The classification for SMOTE data with the first two principal components is shown in Fig. 15.

We applied the PCLFS and RFE feature selection methods to select features from the sample by fitting it on five classification methods. The final ranked feature list for RFE was obtained by considering the feature

Table 5
Feature selection with SMOTE.

Feature	pclfs_order	PCLFS					RFE				
		Logit	LGBM	dt	RFC	SVM-Linear	Logit	LGBM	dt	RFC	SVM-Linear
F22S	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
F21S	2	✓	✓	✓	✓	✓	✓		✓	✓	✓
F21R	3	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
F22R	4	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
F13S	5	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
F15S	6	✓	✓	✓	✓	✓	✓		✓	✓	
F13R	7	✓	✓	✓	✓	✓	✓		✓	✓	✓
F20S	8	✓	✓	✓	✓	✓	✓	✓	✓	✓	
F15R	9	✓	✓	✓	✓	✓	✓		✓	✓	✓
F18S	10	✓	✓		✓	✓		✓	✓	✓	✓
F20R	11	✓	✓		✓	✓			✓	✓	
F5S	12	✓	✓		✓	✓	✓	✓	✓	✓	✓
F9S	13	✓	✓		✓		✓		✓		✓
F1S	14	✓			✓			✓	✓	✓	✓
F5R	15	✓			✓		✓		✓	✓	✓
F4S	16	✓			✓			✓	✓	✓	
F9R	17	✓			✓				✓	✓	✓
F4R	18	✓			✓		✓	✓	✓	✓	✓
F19S	19	✓			✓			✓	✓	✓	
F1R	20	✓			✓		✓	✓	✓	✓	✓
F18R	21	✓			✓		✓		✓		✓
F8S	22	✓			✓		✓	✓	✓	✓	✓
F8R	23	✓			✓			✓	✓	✓	✓
F12S	24	✓			✓		✓	✓	✓		✓
F12R	25				✓		✓		✓		
F14S	26				✓			✓	✓	✓	✓
F19R	27				✓		✓	✓	✓	✓	✓
F2S	28				✓				✓	✓	
F3S	29				✓		✓	✓	✓	✓	✓
F10S	30				✓		✓		✓		
F6S	31				✓		✓	✓	✓	✓	✓
F7R	32				✓		✓	✓	✓	✓	✓
F3R	33				✓		✓		✓	✓	
F16S	34				✓		✓	✓	✓	✓	✓
F7S	35				✓		✓	✓	✓	✓	✓
F14R	36				✓		✓	✓	✓		✓
F6R	37				✓		✓		✓	✓	
F11S	38				✓		✓		✓	✓	✓
F17S	39				✓		✓	✓	✓	✓	
F2R	40				✓		✓		✓	✓	✓
F10R	41				✓		✓		✓	✓	✓
F16R	42				✓		✓	✓	✓		✓
F11R	43						✓		✓		✓
F17R	44						✓	✓	✓	✓	✓

Table 6
Application results for the 5-fold cross-validation.

Model	FS Method	5-fold cross-validation averages			
		F1-score	Precision	Recall	#Features
Logit	basic	0.3792	0.3779	0.3818	44
	rfe	0.3721	0.4307	0.3455	14.2
	pclfs	0.5041	0.6633	0.5091	31.6
LGBM	basic	0.4553	0.5616	0.4182	44
	rfe	0.4443	0.4777	0.4545	27
	pclfs	0.5982	0.7767	0.5818	27
Decision Tree	basic	0.3988	0.3695	0.4364	44
	rfe	0.4248	0.4370	0.4182	12
	pclfs	0.5891	0.7378	0.6727	6
RFC	basic	0.3399	0.6367	0.2364	44
	rfe	0.3468	0.4867	0.2727	6.2
	pclfs	0.5816	1.0000	0.5091	12.2
SVM-Linear	basic	0.4111	0.3917	0.4364	44
	rfe	0.5004	0.4794	0.5273	21.6
	pclfs	0.5958	0.6814	0.6364	25.6

importance of the selected subset and the feature raking of the original classification model. Feature rankings (with SMOTE) relevant to each method and the Bland–Altman (B&A) plots which used to determine the agreement between two pairs of quantitative rankings are presented and discussed in Appendix B.

Table 4 shows the application final results for RFE and PCLFS with different classification models. According to the accuracy measures, F1-score, Precision, and Recall, the PCLFS performs better than the existing RFE method in feature selection.

Then, we selected the relevant feature subset using PCLFS and RFE for each classification model, and the features selected by each method are presented in Table 5.

Finally, We obtained the average cross-validation F1-score, Precision, and Recall for the application data without SMOTE, and the results are presented in Table 6. Even with different cross-validation training sets, the proposed method performs better than RFE with different classification models.

5. Discussion

Feature selection is a crucial phenomenon in high-dimensional classification problems and selecting the informative features in the data set is an essential aspect in feature selection.

Most wrapper feature selection methods use an already ordered feature list as an initial step of selecting features. Still, the question is, how strong and reliable is the ordering ability of the method used to rank the features according to their importance. Also, the order of the

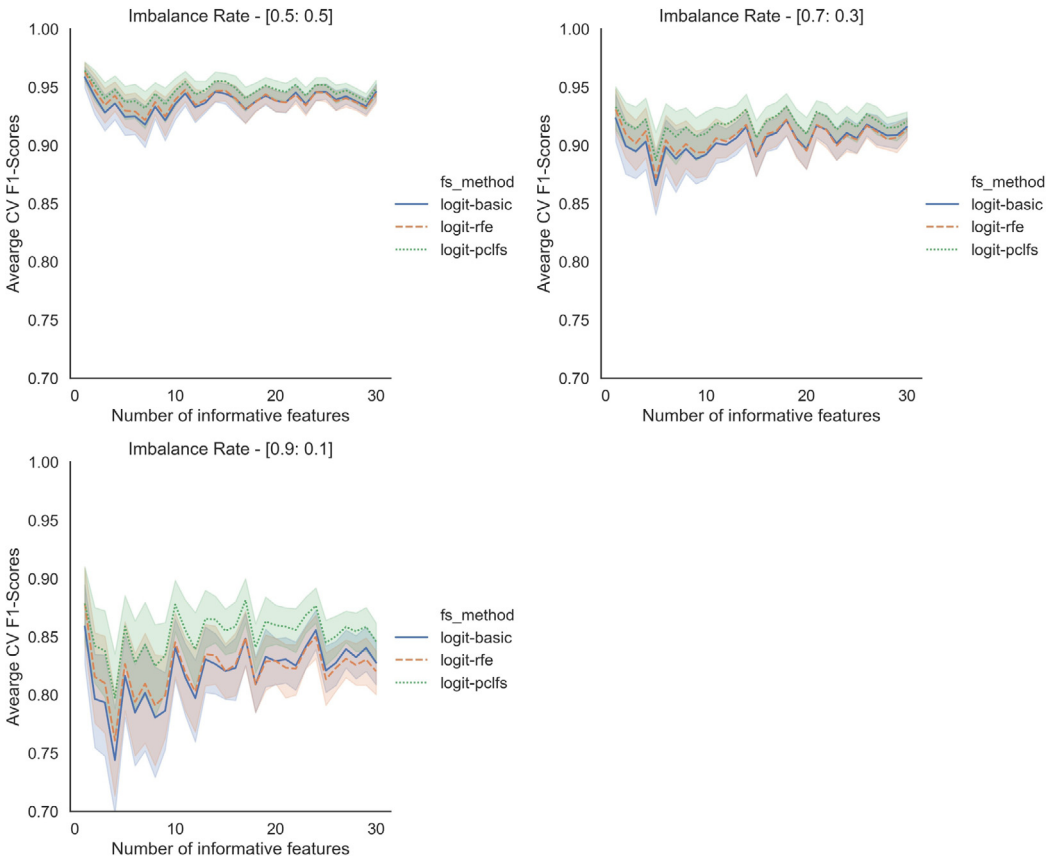

Fig. 13. The averages of the model F1-scores results of a 5-fold cross-validation run.

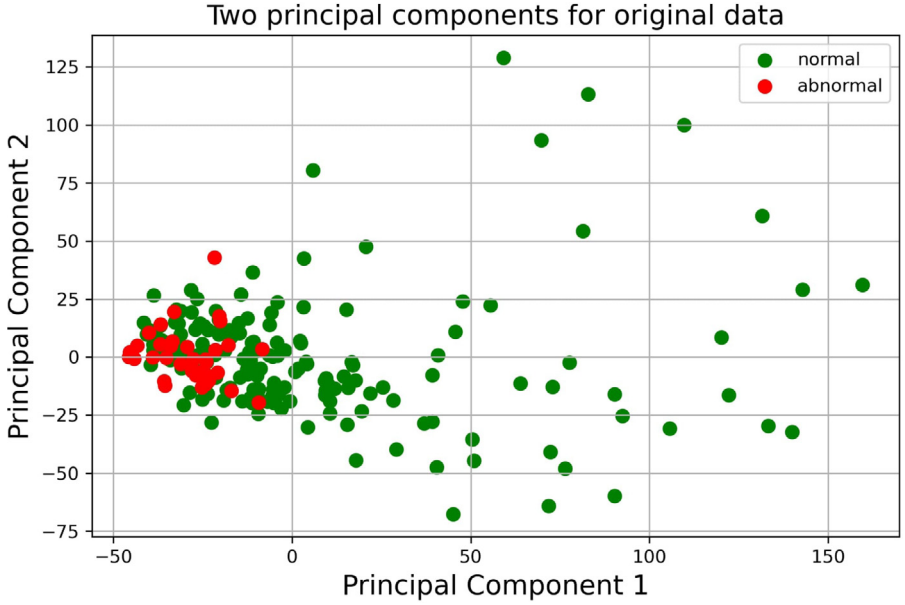

Fig. 14. Classification for first two principal components for original data.

feature set would highly depend on the classification model used in the problem. In this research, we compare four different feature ordering techniques. Using synthetic data, we identify the best feature ordering mechanism as a solution to this issue.

The absolute sum of principal component loadings orders the features more informatively than other selected methods when the number of informative features is not small in the sample. Therefore, we introduced a feature selection method (PCLFS), which uses the absolute sum of principal component loadings to rank the features and a sequential search method to select the feature subset. The PCLFS performs much better with smaller sample sizes and highly imbalanced data sets. The suggested method identifies the most informative features first. Therefore, without applying sequential feature selection, the user can also have any affordable number of features only considering the order of the sum of absolute values of PC loadings.

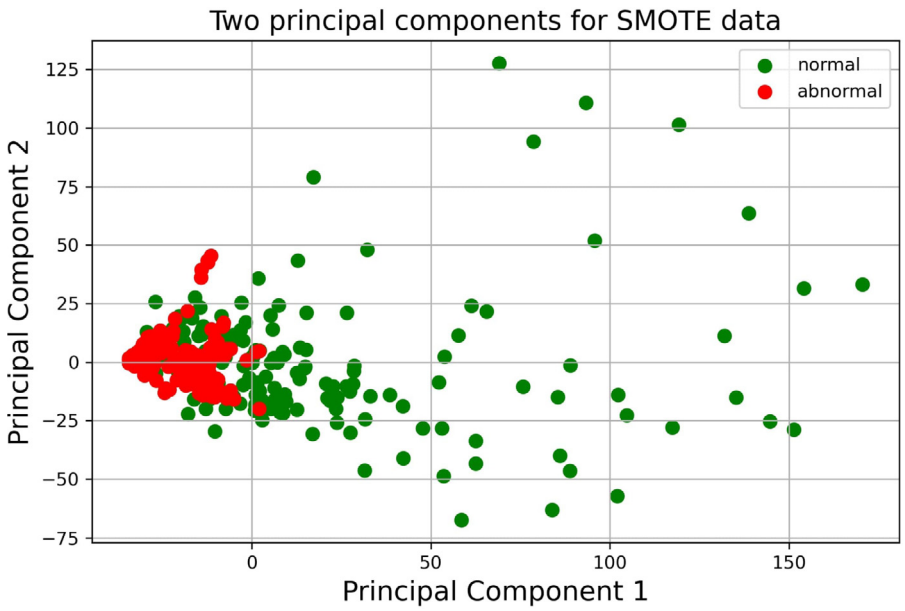

Fig. 15. Classification for first two principal components for SMOTE data.

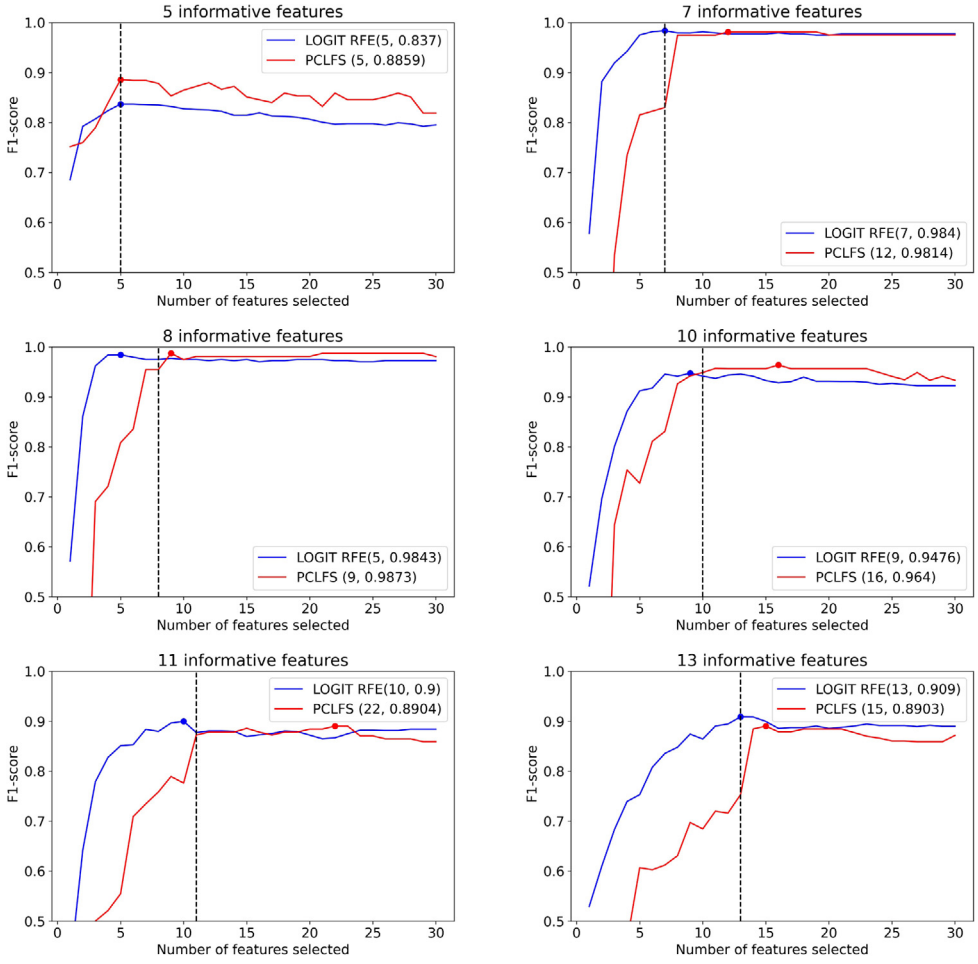

Fig. A.16. Comparison between Logit-RFE cross-validation F1-scores and PCLFS F1-scores with different informative features. The black dashed line indicates the number of informative features in the data set.

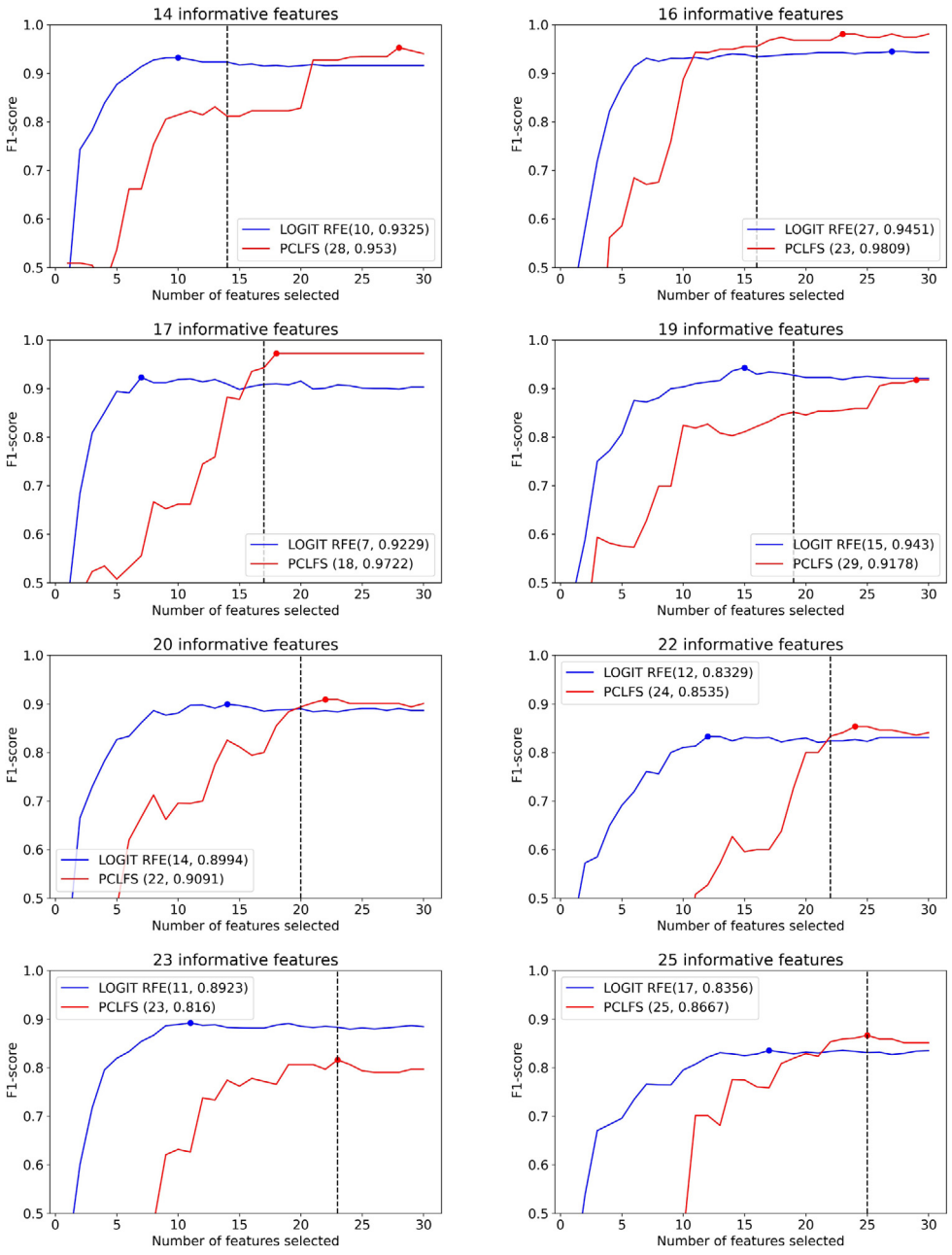

Fig. A.17. Comparison between Logit-RFE cross-validation F1-scores and PCLFS F1-scores with different informative features. The black dashed line indicates the number of informative features in the data set.

The user can select the number of PCs by considering the contribution of each principal component to the total explained variance. The validity of the assumptions and limitations of the PCA (Linearity, Large variances have important structure, and The principal components are orthogonal) are important for the suggested approach (Shlens, 2014). Also, the multiple variables need to be measured at the continuous level and need to have a large enough sample size. Data also must be suitable for data reduction, and there should be no significant outliers in the data set. Although principal components try to cover maximum variance among the features in a data set, if we do not select the number of Principal Components with care, it may miss some information compared to the original list of features. Hence, while we have demonstrated the usefulness of this approach when $k = 2$ principal components are used, the impact of the number of principal components for the suggested method also should be examined in the future.

6. Conclusion

In this study, a feature selection technique is proposed to select the most informative feature subset with better performance. The simulation study has proven the ability of the absolute sum of principal component loadings to order features according to the importance compared to the other method considered. The suggested PCLFS approach of selecting important feature subsets using the absolute sum of principal component loadings performs better than the existing RFE under different conditions. The application results ultimately ensure the accuracy of the findings. However, with a very small number of informative features, the absolute sum of principal component loadings does not order the features well. So, we do not recommend using PCLFS if the number of features that influence the model is believed to be very small.

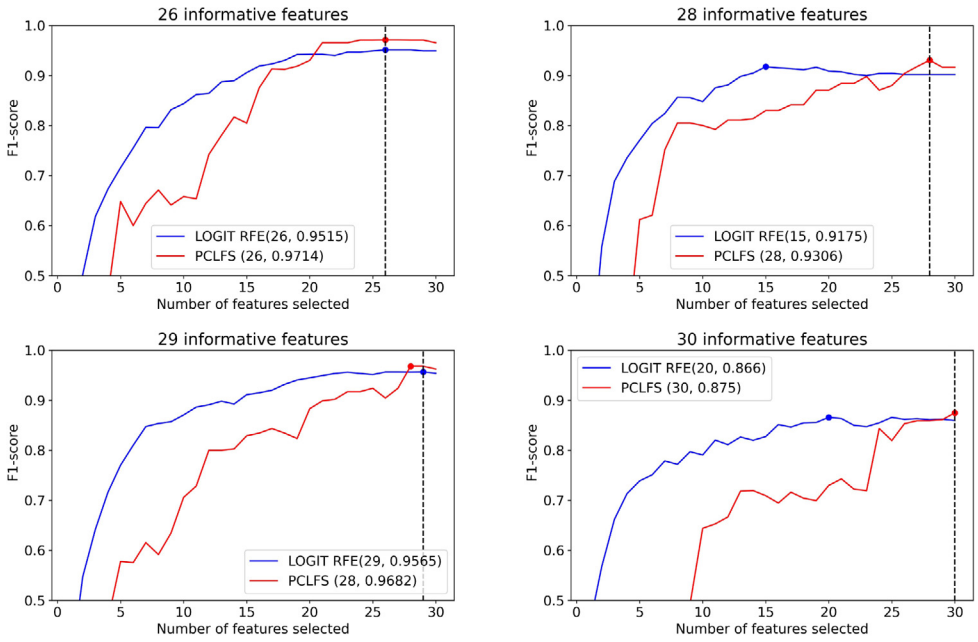

Fig. A.18. Comparison between Logit-RFE cross-validation F1-scores and PCLFS F1-scores with different informative features. The black dashed line indicates the number of informative features in the data set.

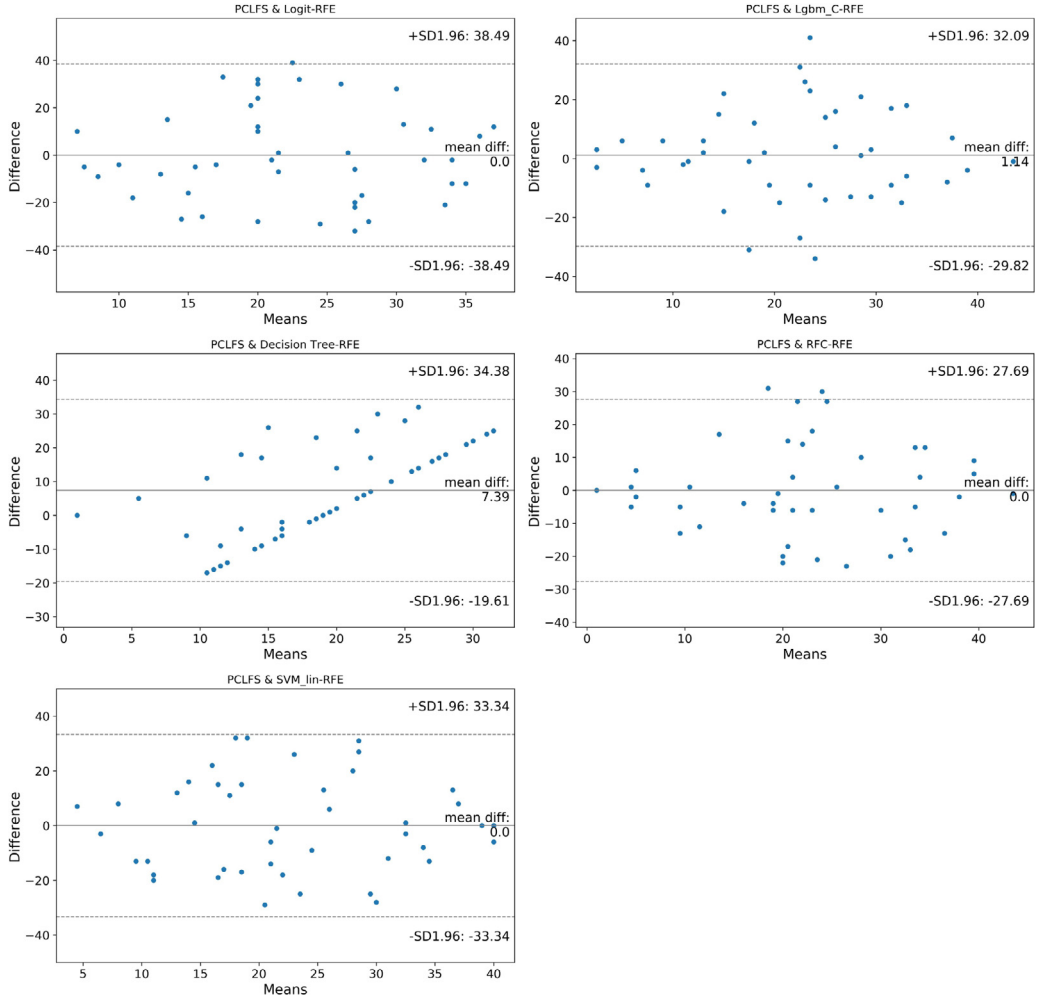

Fig. A.19. Bland and Altman plot for data from Table B.7 by comparing PCLFs with each RFE model, with the representation of the limits of agreement (dotted line), from −1.96s to +1.96s.

Table B.7
Feature ordering ranks for each feature selection method (with SMOTE).

Features	PCLFS	Logit_RFE	Lgbm_RFE	DT_RFE	RFC_RFE	SVM_RFE
F22S	1	28	4	1	1	21
F21S	2	20	33	19	7	20
F21R	3	29	12	19	16	16
F22R	4	13	1	19	6	17
F13S	5	10	9	19	4	8
F15S	6	34	24	12	17	35
F13R	7	23	41	16	12	26
F20S	8	12	2	3	2	1
F15R	9	17	36	19	31	25
F18S	10	39	12	19	30	27
F20R	11	43	12	15	10	36
F5S	12	2	6	19	29	4
F9S	13	18	28	19	34	31
F1S	14	42	12	18	18	28
F5R	15	19	24	17	38	14
F4S	16	38	10	5	22	44
F9R	17	37	18	19	21	42
F4R	18	25	32	19	24	24
F19S	19	36	28	19	20	7
F1R	20	22	18	19	26	29
F18R	21	6	34	19	41	22
F8S	22	21	7	4	5	6
F8R	23	44	36	6	19	12
F12S	24	30	12	19	42	9
F12R	25	15	40	19	40	37
F14S	26	14	4	19	25	11
F19R	27	26	36	13	33	5
F2S	28	40	24	2	13	41
F3S	29	41	28	19	15	23
F10S	30	9	36	7	43	38
F6S	31	33	28	14	36	34
F7R	32	8	18	19	14	19
F3R	33	35	41	19	23	32
F16S	34	1	18	9	3	2
F7S	35	5	12	19	8	3
F14R	36	4	10	19	32	10
F6R	37	24	41	19	39	43
F11S	38	27	7	8	11	18
F17S	39	7	18	11	9	39
F2R	40	32	23	19	27	40
F10R	41	11	34	19	28	33
F16R	42	3	24	10	37	15
F11R	43	31	44	19	44	30
F17R	44	16	3	19	35	13

CRediT authorship contribution statement

Surani Matharaarachchi: Investigation, Methodology, Formal analysis, Data curation, Software, Validation, Visualization, Writing – original draft. **Mike Domaratzki:** Conceptualization, Supervision, Writing – review & editing. **Saman Muthukumarana:** Conceptualization, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Muthukumarana and Domaratzki have been partially supported by research grants from the Natural Sciences and Engineering Research Council of Canada.

Appendix A. F1-score comparison with different informative features

Figs. A.16, A.17, and A.18 present similar results as Fig. 9 but for different number of informative features in the data set.

Appendix B. Feature ranking (with SMOTE)

Feature rankings (with SMOTE) relevant to each method considered are shown in Table B.7. Then the Bland–Altman (B&A) plots (Bland & Altman, 1999) were used to determine the agreement between two pairs of quantitative rankings. The method quantifies agreement between two quantitative measurements by constructing limits of agreement for each RFE model with PCLFS. These statistical limits were calculated using the mean and the standard deviation (s) of the differences between the two ranks. First, we checked the assumptions of normality of differences using the Shapiro normality test, and all differences were concluded as normally distributed. Since the points on the plots (Fig. A.19) are scattered above and below zero and within the limits, it suggests no consistent bias of one approach versus the other. The straight line on the decision tree-RFE plot and the results from Table 4 indicate that for the decision tree-RFE has given equal ranking for several features while PCLFS ranks them differently.

References

Bellman, R. (1957). *Dynamic programming*. Princeton: Princeton University Press.
Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, *8*(2), 135–160.
Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1).
Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: JWadsworth & Brooks/Cole Advanced Books & Software.
Bruyant, P. P. (2002). Analytic and iterative reconstruction algorithms in SPECT. *Journal of Nuclear Medicine*, *43*(10), 1343–1358.
Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *The Journal of Artificial Intelligence Research*, *16*, 321–357.
Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.
Dunteman, G. (1989). Using principal components to select a subset of variables. In *Quantitative applications in the social sciences*, *Principal components analysis*. Newbury Park: SAGE Publications, Inc.
Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232.
Guo, Q., Wu, W., Massart, D. L., Boucon, C., & De Jong, S. (2002). Feature selection in principal component analysis of analytical data. *Chemometrics and Intelligent Laboratory Systems*, *61*(1–2), 123–132.
Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1), 389–422.
Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *Springer series in statistics*, *The elements of statistical learning : Data mining, inference, and prediction* (2nd ed.). New York: Springer.
Hotelling, H. (1933). *Analysis of a complex of statistical variables into principal components*. Baltimore: Warwick & York.
Krzysztof, C. J., Daniel, W. K., & Ning, L. (1997). CLIP3: Cover learning using integer programming. *Kybernetes*, *26*(5).
Kuhn, M. (2013). *Applied predictive modeling* (1st ed.). New York, NY: Springer New York.
Kumari, B., & Swarnkar, T. (2011). Filter versus wrapper feature subset selection in large dimensionality micro array: A review. *International Journal of Computer Science and Information Technologies*, *2*, 1048–1053.
Kurgan, L. A., Cios, K. J., Tadeusiewicz, R., Ogiela, M., & Goodenday, L. S. (2001). Knowledge discovery approach to automated cardiac SPECT diagnosis. *Artificial Intelligence in Medicine*, *23*(2), 149–169.
Lal, T. N., Chapelle, O., Weston, J., & Elisseeff, A. (2006). Embedded methods. In *Studies in fuzziness and soft computing*, *Feature extraction* (pp. 137–165). Berlin, Heidelberg: Springer.
Lev, J. (1949). The point biserial coefficient of correlation. *The Annals of Mathematical Statistics*, *20*(1), 125–126.
McCullagh, P., & Nelder, J. A. (1989). *Monographs on statistics and applied probability*: *vol. 37*, *Generalized linear models* (2nd ed.). London: Chapman & Hall.
Miche, Y., Bas, P., Lendasse, A., Jutten, C., & Simula, O. (2007). Advantages of using feature selection techniques on steganalysis schemes. In F. Sandoval, A. Prieto, J. Cabestany, & M. Graña (Eds.), *Computational and ambient intelligence* (pp. 606–613). Berlin, Heidelberg: Springer.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peng, Y., Wu, Z., & Jiang, J. (2010). A novel feature selection approach for biomedical data classification. *Journal of Biomedical Informatics*, *43*(1), 15–23.

Pudil, P., Novovičová, J., & Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, *15*(11), 1119–1125.

Saeys, Y., Inza, I. n., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517.

Shlens, J. (2014). A tutorial on principal component analysis.

Stańczyk, U. (2015). Feature evaluation by filter, wrapper, and embedded approaches. In *Feature selection for data and pattern recognition* (pp. 29–44). Berlin, Heidelberg: Springer.

Tate, R. F. (1954). Correlation between a discrete and a continuous variable. point-biserial correlation. *The Annals of Mathematical Statistics*, *25*(3), 603–607.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B, Methodological*, *58*(1), 267–288.

Tsuruoka, Y., Tsujii, J., & Ananiadou, S. (2009). Stochastic gradient descent training for L1-regularized log-linear models with cumulative penalty. In *ACL '09*: *vol. 1*, *Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP* (pp. 477–485). USA: Association for Computational Linguistics.

Weisberg, S. (2005). *Applied linear regression*. Hoboken: John Wiley & Sons, Incorporated.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83.

Xia, G., & Jin, W. (2008). Model of customer churn prediction on support vector machine. *Systems Engineering - Theory & Practice*, *28*(1), 71–77.