Feature Selection by mRMR Method for Heart Disease Diagnosis

GAOSHUAI WANG1, FABRICE LAURI2, AND AMIR HAJJAM EL HASSANI1
1Laboratoire Nanomédicine, Imagerie, Therapetique-EA4662, Université Bourgogne Franche-Comté, UTBM, 90010 Belfort, France
2Connaissance et Intelligence Artificielle Distribuées (CIAD)-UMR 7533, Université Bourgogne Franche-Comté, UTBM, 90010 Belfort, France
Corresponding author: Gaoshuai Wang (gaoshuai.wang@utbm.fr)

This work was supported in part by the China Scholarship Council (CSC) under Grant CSC N° 201806280502.

ABSTRACT Heart disease has become a non-ignorable threat to human health in recent years. Once without timely diagnosis and treatment, patients often suffer disability or even death. However, the diagnosis accuracy directly relies on different doctors’ experiences and various factors associated with heart disease bring heavy tasks on them make the situation worse. Therefore, to improve heart disease treatment, introducing computer-aided techniques to assist doctors in diagnosis is a feasible approach. At present, researchers usually use the processed dataset (13 features) selected by doctors from the unprocessed dataset (74 features) (UCI Machine Learning Repository) and apply the feature selection method to the dataset, it’s inappropriate because the feature scale is so small. People neglect the unprocessed dataset’s value and don’t realize it could contain some latent information. A comprehensive comparison is needed to demonstrate the unprocessed dataset’s advantages. Besides, the incremental feature combination method should be verified. As the minimum Redundancy - Maximum Relevance (mRMR) gains great success in feature selection, applying it as a feature filter can enhance classification accuracy. Thus, in this research, we introduced the mRMR method as a filter for feature selection and made a comprehensive comparison within several methods like Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Kendall, Random Forest, and other research works in several metrics. By analyzing the results, in most cases, the unprocessed dataset can enhance algorithm’s performance. The incremental feature selection method is effective and the mRMR is superior to other methods. Not only does it own the highest accuracies, but also the least supportive features. It has 100% accuracy with 8 features on the Cleveland dataset, 98.3% accuracy with 14 features on Hungarian, and 99% accuracy with 9 features on Long-beach-VA, respectively. Furthermore, we find that some features, which doctors regard as useless, play a part in classification, that should attract some attention from doctors.

INDEX TERMS Heart disease, feature selection, mutual information, mRMR.

I. INTRODUCTION
A. BACKGROUND AND CURRENT STUDY SITUATION
Cardiovascular diseases (CVDs) causes millions of death every year due to their suddenness. Once without the treatment timely, the patient is prone to suffer disability or even death. Therefore, early and accurately diagnosing heart disease is an effective way to save lives [1].

With the development of medical science, doctors have discovered lots of symptoms associated with heart disease, such as high blood pressure, stress, diabetes, etc [2], [3], [4]. But, these factors only indicate that the patient has a high possibility of suffering from heart disease. The diagnosis accuracy depended on the doctor’s experience will lead to misdiagnosis and missed diagnoses. To overcome the deficiency in current situation and the prosperity of AI [5], [6], researchers have tried to make use of AI’s power to assist doctors in medical diagnosis [7], [8]. In some areas, there is a reality that AI’s diagnosis ability has suppressed human beings [9].

In the beginning, some researchers only used the classifiers to deal with the original dataset. Atallah et al. [10]
directly adopted the majority voting method to produce a more powerful classifier. This ensemble method consists of four machine learning methods, including Stochastic Gradient Descent (SGD), K-Nearest Neighbor, Random Forest, and Logistic Regression. The highest prediction accuracy is 88% among the four algorithms. By using the voting technique, the ensemble method achieves 90% accuracy. These kinds of methods ignore the feature’s difference and make no full use of the dataset.

Reducing dataset’s dimensionality, which means eliminating redundant or irrelevant features [11] that will significantly degrade the performance of machine learning methods or increase computation costs, is a necessary step in current research area. Dimensionality reduction has two main branches, feature extraction, and feature selection. PCA and LDA are the typical feature extraction methods [12]. Karayilan et al. [13] put up a neural network method with PCA to recognize heart disease. Kolukisa et al. [14] also proposed an ensemble algorithm based on the LDA feature selection. Shafizadeh-Moghadam and Hossein [15] put up a method for selecting effective components in PCA and as the input of Random Forest on spectroscopic data. They found that the most relevant principal components to the target variable were not the first few. However, these methods will lose the interpretation from the original dataset.

Feature selection picks up a subset of features obeying a certain algorithm from the original dataset. These methods reserve all feature information and make prediction results more interpretable. Further, it’s also a powerful tool to prevent overfitting [16] caused by the increment of feature dimensionality space. Generally, there exist three categories of feature selection, filter methods, wrapper methods, and embedded methods [17]. Filter methods use some criteria to evaluate their rank or importance. It’s computationally efficient and can obtain a pretty good result in a short time; Enriko et al. [18] applied the Chi-Squared to select key features and give the 6 top variables a more important weight which equals 2, the other features’ weight is 1. Lutu et al. [19] put up a new definition to evaluate the correlation between features, they classified the correlation into three categories, high, medium, and low. Then, using the heuristic algorithm to search for the optimal combination. Results show that the feature selection method provides a high level of predictive classification performance. You et al. [20] reduced the high-dimensional multi-category data by the PLS-based local recursive feature elimination. They considered that the single-feature measure methods to evaluate the importance of feature is based on the assumption of independence among features features’ correlation will influence the prediction. But, these methods ignore the interaction within features.

Wrapper methods incorporate a classifier into a pre-determined objective function that evaluates the appropriateness of the predictor subsets through an exhaustive search. The optimization of feature combinations is an NP-hard problem. To obtain the optimal feature combination, all $2^N - 1$ combinations should be tested by classifiers, assuming the feature number is $N$, and the time complexity will increase exponentially [21]. Amini et al. [22] used a two-layer feature selection method to resolve the regression problem. This method consists of a genetic algorithm and elastic net. The genetic algorithm, one of the heuristic methods is the first layer to select a local optimization feature subset, the elastic net as the second layer eliminates the redundant features by using the penalty factors. Based on current computation ability, it’s unrealistic to search for the optimal feature combination [23].

Since the emergence of Shannon information entropy [24], filter methods based on this theory play an important role in AI, such as the famous decision tree algorithm. Mutual information maximization (MIM) method is first put up for reducing class labels’ uncertainty. However, this method merely considers a feature’s relevancy and ignores its redundancy, which makes the redundancy exist in the selected features. To promote mutual information’s effect, Peng et al. [25] introduce the conception of feature redundancy into mutual information, which greatly improves its applicability. In heart disease datasets, value missing is a common phenomenon and has a profound negative influence on feature selection methods. The mRMR method will ignore specific feature values and resist this negative influence brought by missing values. Yan et al. [26] employed the mRMR method for selecting the candidate feature for the fault diagnosis of rotating machinery. Özyurt and Faith [27] diagnosed the white blood cell test by the fused CNN model with mRMR feature selection. Eroglu et al. [28] proposed the mRMR-based hybrid CNN model for the classification of Alzheimer’s disease. Narender and Dharmender [29] used the mRMR as the feature filter in heart disease diagnosis on ECG dataset. Various applications of mutual information methods indicate mRMR has a strong generalization and its ability is extensively verified.

B. OUR CONTRIBUTIONS

Current works still have these drawbacks: they mainly apply some algorithms and feature selection methods to the 13 features dataset. We think that there is not necessary to employ the feature selection method on such a few number features. The algorithm’s performance is limited by the dataset, and the dataset with 74 features should contain more supportive information. The interaction within features must be considered, the mutual information method is a good choice. Besides, the incremental feature combination method’s effectiveness should be verified. In this research, we emphasized the importance of 74 features dataset and removed the performance limitation from the dataset. The mRMR method is selected as the feature selection method, it’s one of the filter methods and has high efficiency, well considering the interaction within features. What’s more, we verified the effectiveness of the incremental feature combination method and conducted comprehensive experiments.
Along with the popular works, to receive better prediction results and reduce the classifier characteristic’s influence, six most used classifiers such as Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF), Linear Regression (LR), Gaussian Bayesian (GB), and K-nearest neighbors (KNN) are applied in research.

The rest of this paper is organized as follows. Section II presents theories and algorithms of PCA, LDA, Kendall $\tau$ correlation, Random Forest algorithm, and mRMR. Besides, this part also introduces heart disease datasets and the flowchart of the whole experiment. Section III displays a comprehensive comparison of results, discussion, and conclusion. Finally, Section IV summaries all results in this work and puts up a further plan.

II. METHODS AND DATASETS

A. METHODS

Dimensionality reduction [30] is a technique to reduce feature scale. It not only reduces their dimensionality and accelerates the training process, but also discovers the latent information. Sometimes, this kind of method can promote some features’ importance [31], [32]. This part will represent the details of PCA, LDA, Kendall $\tau$ correlation, Random Forest, and mRMR methods.

1) PCA METHOD

PCA is a kind of method which explores the most significant factors and represents the original data. It uses matrix transformation to achieve the goal of dimensionality reduction, makes information loss as small as possible, and lets the values of variances become as large as possible. PCA has been used in many research areas, like image recognition, and text mining. Let $W$ represent the transformation matrix that transfers the original f-dimensional feature matrix $X$ into the f-dimensional feature matrix $Y$ where $f \leq t$. The new feature vector $y_i = W^T x_i$, $i = 1, \ldots, N$. The columns of $W$ are the eigenvalues $\lambda_i e_i$ obtained by solving the eigenstructure decomposition $\lambda_i e_i = Q e_i$, where $Q = XX^T$ is the covariance matrix and $\lambda_i$ is the eigenvalue associated with the eigenvector $e_i$.

2) LDA METHOD

LDA is a transformation that yields the largest differences inter-class and smallest differences intra-class. There are two measures: within-class scatter matrix $S_w$ and $S_b$ between-class scatter matrix.

$$\begin{align*}
S_w &= \sum_{j=1}^{c} N_j \sum_{i=1}^{N_j} (X_i^j - \mu_j)(X_i^j - \mu_j)^T \\
S_b &= \sum_{j=1}^{c} (\mu_j - \mu)(\mu_j - \mu)^T 
\end{align*}$$

where $\mu_j$ is the mean of all classes. To achieve the aim of the largest inter-class differences and the smallest intra-class differences, we can maximize the ratio $\dfrac{\det[S_b]}{\det[S_w]}$. If $S_w$ is a nonsingular matrix, this ratio is maximized when the column vector of the projection matrix, $W$, are the eigenvectors of $S_w^{-1}S_b$.

3) KENDALL CORRELATION

Kendall’s $\tau$ correlation is an index to make sure whether there is a correlation between two variables. It’s defined as the difference between the number of pairs of concordant and discordant values, normalized by the total number of pairs. Let $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ be two sequences to calculate. Define $A$ as the set of all pairs of indices:

$$A = \{(i, j) \in [1, \ldots, n] \times [1, \ldots, n] | i < j\}$$

and let $n_A = \frac{1}{2}n(n-1)$. $C \subseteq A$, with cardinal $n_C$; $D \subseteq A$, with cardinal $n_D$.

$$C = \{(i, j) \in A | (x_i < x_j \text{ and } y_i < y_j) \text{ or } (x_i > x_j \text{ and } y_i > y_j)\}$$

$$D = \{(i, j) \in A | (x_i < x_j \text{ and } y_i > y_j) \text{ or } (x_i > x_j \text{ and } y_i < y_j)\}$$

Kendall’s $\tau$ rank correlation is defined as:

$$\tau(X, Y) = \frac{n_C - n_D}{n_A}$$

The value of $\tau$ ranges from 1 when $n_C = n_A$, to -1, $n_D = n_A$.

Sometimes, for eliminating the influence of the pairs when $x_i = x_j$ or $y_i = y_j$, Kendall’s correlation will be change to define as:

$$\tau(X, Y) = \frac{n_C - n_D}{n_C + n_D}$$

Many factors linked with heart disease appear in one person and their values are high always means the person has a high potential risk of heart disease. Besides, in the dataset, label 0 represents health, and label 1 means illness, smartly, treating label values as the extent of the disease satisfies Kendall’s application condition. Due to Kendall’s correlation method just judges the connection strength between two variables, thus, it can handle the interference of the non-linear correlation. As we know, no researcher introduces Kendall correlation into heart disease prediction. The detailed algorithm of Kendall correlation feature selection is described in algorithm 15. We defined a Rule for eliminating some features directly, like missing too many values.

4) RANDOM FOREST

Breiman [33] invented the decision tree algorithm in the 1980s. In 2001, he assembled the decision trees into the Random Forest by randomly using the variables to generate many decision trees. The Random Forest algorithm is as follows:

1) Randomly choosing many subsets from the training set $S$, get a series of subsets $S_1, \ldots, S_n$. 

2) Using the subsets to train the decision trees. In the training process, the split rule for each node is randomly
Algorithm 1: Kendall τ

Input: The heart disease dataset \( D \), selected feature number \( n \);
Output: The heart disease dataset \( D_p \) with the target feature;

1: load the original dataset \( D \);
2: extract label \( L \) from dataset \( D \), \( D_y = \emptyset \);
3: for each feature \( f \) in \( D \) do
4: \( D_y \leftarrow D_y \cup f \);
5: end if
6: end for
7: for each feature \( f \) in \( D_y \) do
8: calculate \( \tau_f \) Kendall(f, \( L \));
9: add \( (h_f, \tau_f) \) into dict \( d \);
10: end for
11: sort \( d \) by \( \tau_f \) in descending order;
12: for each \( i \) in \([0, n] \) do
13: \( D_p \leftarrow D_p \cup D_t[d(h_f)] \);
14: end for
15: return \( D_p \);

Selecting \( k \) features from all features, then selecting the optimal split node from these \( k \) features to divide the sub-trees. Repeatedly generating the decision tree.

3) Summarizing the results from each decision tree’s prediction, selecting the final results in the summarized results.

Our feature selection via Random Forest is displayed in algorithm 2. To accelerate the processing speed and dismiss the complexity of building the decision trees in a Random Forest, we categorize the feature which owns more than 20 values into 11 integers, \( var \in [0, 10] \).

5) mRMR

The mRMR considers both feature relevancy with class label and feature redundancy among the selected features. Its formula is as follows:

\[
J(x_k) = I(x_k; y) - \frac{1}{|S|} \sum_{x_j \in S} I(x_j; x_k)
\]

(1)

where \( J(x_k) \) is the evaluation index, \( x_k \) is the candidate feature and \( S \) is the subset of the selected features. The first selected feature is chosen by MIM method. After, we will move one feature with the largest \( J(x_k) \) into the subset of selected features in each update.

Six popular classifiers are used to predict heart disease because there is a performance gap due to classifiers’ characteristics. They are Random Forest (RF), Logistic Regression (LR), Neural Network (ANN), Gaussian Bayes (GB), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM).

- Each dimensionality reduction method will be combined with six classifiers on three heart disease datasets.
- The number of the selected feature is from 5 to 30.
- Training the model with 10-fold cross-validation.

The flowchart of this research is shown in Fig.1. There are two versions of the datasets. After pre-processing, the features will be selected by PCA, Kendall, Random Forest, and mRMR methods, then, the subset of features is sent as the input of classifiers. During the prediction, we adopt the incremental feature combination for pursuing optimal results. The mRMR implements a paired two-tailed t-test with other methods, P-Value is set as 5%. ‘+’, ‘−’ and ‘≈’ indicate that mRMR performs ‘better than’, ‘worse than’, and ‘equal to’ the corresponding method.

B. DATASETS

The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by many machine learning communities for empirical analysis [34]. It provides more than 500 datasets concerning various domains. The heart disease dataset in this research

Algorithm 2: Random Forest

Input: The heart disease dataset \( D \), selected feature number \( n \);
Output: The heart disease dataset \( D_p \) with the target feature;

1: load the original dataset \( D \);
2: extract label \( L \) from dataset \( D \), \( D_y = \emptyset \);
3: for \( i = 1 \) to \( 2 \) do
4: if \( f \neq Rule \) then
5: \( D_y \leftarrow D_y \cup f \);
6: end if
7: end for
8: for each feature \( f \) in \( D_y \) do
9: \( D_y \leftarrow D_y \cup D_t[d(h_f)] \);
10: end for
11: return \( D_p \);

Algorithm 3: mRMR

Input: The training dataset \( D \) with original feature set \( F = \{f_1, f_2, \ldots, f_n\} \);
Output: The selected subset feature \( S_T \);

1: \( S_T \leftarrow \emptyset; \)
2: for \( f_i \) in \( F \) do
3: \( M_{f_i} = I(f_i; y) \);
4: \( f \leftarrow \max(M_f) \);
5: \( S_T \cup f \);
6: end for
7: for \( i \) to \( 2 \) to \( T \) do
8: for each \( f_i \) in \( F \) do
9: \( J(f_i) = M_{f_i} - \frac{1}{|S|} \sum_{f_j \in S} I(f_j; f_i) \);
10: end for
11: select \( f_i \) from \( J(f_i) \) with the largest value;
12: \( S_T = S_T \cup f_i \);
13: \( F = F - f_i \);
14: end for
15: return \( S_T \)
is from the UCI Repository. It’s from four medical agencies, the Hungarian Institute of Cardiology, Zurich University Hospital, Basel University Hospital, and V.A Medical Center.

The heart disease dataset contains 74 features with a detailed description, but the medics only recommended 13 features within them. Thus, there are two versions of datasets in heart disease diagnosis, the unprocessed version (74 features) and the processed version (13 features).

Table 1 displays its detailed description. Particularly, we find that most of the related heart disease factors have a characteristic in their values, smaller factor’s value often demonstrates the patient is prone to be healthy. On the contrary, it is not.

Table 2 shows Cleveland, Hungarian, and Long-Beach-Va’s distributions. They have 283, 294, and 123 samples respectively after eliminating useless or missing values. The absence denotes (health), and the presence denotes (illness). Two of them are imbalanced.

III. RESULTS AND ANALYSIS

A. RESULTS ON THE ORIGINAL DATASET

The results with the original dataset are the benchmark in this research. In Table 3, we compare the results of processed dataset (13 features) and original dataset (74 features) without dimensionality reduction. The characteristics of both classifiers and datasets have a certain influence on the prediction results.

Generally, the original dataset’s results have a better performance, confirming our assumption that the dataset with more original features should outperform the processed dataset. Within these classifiers, ANN and KNN perform worst, nearly losing the prediction function in their respective AUC. It indirectly indicates some classifiers have a strong ability to eliminate irrelevant features, and pre-processing is necessary when applying ANN and KNN. Another notable point is that dataset type has a certain influence on the prediction effect, it seems that Cleveland and Hungarian are better than Long-Beach-Va. We reckon that there exist plenty of missing values in the Long-Beach-Va, these missing values may play a vital role in diagnosing heart disease.

B. THE INFLUENCE OF IMPORTANT FEATURES

In Fig. 2, we test the incremental feature combination method’s performance based on the Kendell 𝜏 and Random Forest methods on three datasets. In this experiment, 30 selected features are accumulated in order, at least number of the initial features is five, the more important features are added later, so the first feature we add is 25th important, then 24th important, and so on. By observing the results, with the increment of feature number, the accuracy of Cleveland and Hungarian grows fast. There is no obvious growth of accuracy on the Long-Beach-Va dataset until the feature number is above 25. Especially, when the feature number is below 15, the accuracy growth is small and stable. It indicates that unimportant features have little influence on the prediction and most of the accuracy promotion is brought by the important features. When the feature number approaches 30 on Cleveland and Long-Beach-Va datasets, each feature addition will bring the obvious accuracy promotion. While this phenomenon nearly disappears from the Hungarian dataset. By analyzing the results, we can conclude that the incremental combination of important features is reasonable and effective for getting a comparatively optimal feature combination. By the way, in the incremental combination methods, the more important features are added firstly, and there are five initial features at the beginning.

C. RESULTS ON CLEVELAND

Fig. 3 represents four feature selection methods’ results on Cleveland dataset. We can see that results of RF, Kendall, and mRMR are similar. By reaching the highest accuracy, RF and Kendall’s results have a ladder growth, which denotes the addition of some features is useless and there exists some redundancy in features; While mRMR’s accuracy grows rapidly, which indicates it can accurately catch the key features. Thus, we can have a conclusion that mRMR method has the best performance because it not only owns the highest result but also the smallest feature number (8). PCA performs worst in the four feature selectors. As for the specific classifiers, KNN has the same tendency in four figures which grows in the beginning and then decreases with the increment of features. Except for KNN, nearly all methods maintain high accuracy.

D. RESULTS ON HUNGARIAN

In Fig. 4, it’s shown that all the classifiers’ results have no obvious growth except Random Forest. It denotes that most of the selected features are redundant and Random Forest has the advantage of mining the latent information compared with other classifiers. Besides, we can discover that nearly all the accuracies are low than the results on Cleveland, which validates previous experimental results, the phenomenon of KNN’s accuracy decline also disappears in the Hungarian dataset. For accuracy, mRMR should have the best performance, RF achieves the highest accuracy, more than 98%.
TABLE 1. Features of the processed datasets (Cleveland, Hungarian, and Long-Beach-Va).

| Serial Number | Original Number | Feature Name | Feature Description |
|---------------|-----------------|--------------|---------------------|
| 1             | 3               | age          | age in years        |
| 2             | 4               | sex          | sex (1 = male; 0 = female) |
| 3             | 9               | cp           | cp: chest pain type – Value 1: typical angina– Value 2: atypical angina– Value 3: non-anginal pain– Value 4: asymptomatic |
| 4             | 10              | thumbps     | resting blood pressure |
| 5             | 12              | chol         | serum cholesterol in mg/dl |
| 6             | 16              | fbs          | (fasting blood sugar >120 mg/dl) (1 = true; 0 = false) |
| 7             | 19              | restecg     | resting electrocardiographic results |
| 8             | 32              | thalach     | maximum heart rate achieved |
| 9             | 38              | exang       | exercise induced angina (1 = yes; 0 = no) |
| 10            | 40              | oldpeak     | ST depression induced by exercise relative to rest |
| 11            | 41              | slope       | the slope of the peak exercise ST segment |
| 12            | 44              | ca          | height at rest |
| 13            | 51              | thal        | 3 = normal; 6 = fixed defect; 7 = reversible defect |
| 14            | 58              | num         | diagnosis of heart disease (angiographic disease status) |

TABLE 2. The processed and original datasets distribution.

| Dataset             | Total instances | Presence | Absence |
|---------------------|-----------------|----------|---------|
| Cleveland           | 283             | 157 (55%)| 126 (45%)|
| Hungarian           | 294             | 188 (64.9%)| 106 (35.1%)|
| Long-Beach-Va       | 123             | 93 (75.6%)| 30 (24.4%)|

TABLE 3. The results of two versions datasets, results of original datasets are in parentheses.

| Dataset             | Method | Accuracy(%) | Precision(%) | Recall(%) | F1(%) | AUC(%) |
|---------------------|--------|-------------|--------------|----------|-------|--------|
| Cleveland           | RF     | 81.5 (94.6) | 81.2 (97.2)  | 75.7 (90.2) | 78.0 (93.5) | 81.0 (94.1) |
|                     | LR     | 83.2 (92.6) | 86.6 (97.4)  | 77.2 (85.5) | 81.4 (90.9) | 83.0 (91.9) |
|                     | ANN    | 61.4 (57.1) | 58.2 (46.8)  | 68.1 (62.0) | 59.4 (49.6) | 62.4 (58.2) |
|                     | GB     | 83.7 (98.8) | 84.8 (97.8)  | 79.1 (99.4) | 81.6 (98.6) | 83.4 (98.8) |
|                     | KNN    | 63.8 (63.1) | 59.4 (60.9)  | 57.9 (49.9) | 58.4 (54.5) | 63.2 (61.9) |
|                     | SVM    | 85.0 (89.5) | 85.5 (91.0)  | 82.0 (84.0) | 83.6 (87.2) | 84.8 (88.9) |
| Hungarian           | RF     | 80.9 (92.6) | 75.9 (91.5)  | 69.6 (88.2) | 72.2 (89.7) | 75.5 (91.7) |
|                     | LR     | 83.0 (91.0) | 78.4 (88.6)  | 69.6 (86.2) | 73.4 (87.1) | 79.8 (90.0) |
|                     | ANN    | 63.4 (63.3) | 11.3 (3.5)   | 15.6 (0.8)  | 13.1 (1.3)  | 52.8 (49.9) |
|                     | GB     | 82.6 (84.8) | 76.4 (76.5)  | 79.5 (84.8) | 77.7 (79.8) | 82.1 (84.7) |
|                     | KNN    | 64.0 (68.4) | 52.5 (58.4)  | 58.8 (39.6) | 44.1 (45.6) | 58.8 (61.0) |
|                     | SVM    | 80.9 (89.6) | 73.7 (85.3)  | 72.7 (86.1) | 72.7 (85.5) | 79.0 (88.8) |
| Long-Beach-Va       | RF     | 76.2 (87.3) | 79.7 (87.9)  | 91.8 (97.1) | 85.2 (92.1) | 60.8 (77.0) |
|                     | LR     | 74.0 (82.5) | 75.7 (88.3)  | 95.8 (88.7) | 84.4 (88.4) | 54.0 (76.2) |
|                     | ANN    | 71.4 (68.5) | 73.4 (77.1)  | 95.5 (82.8) | 82.8 (79.6) | 51.6 (54.8) |
|                     | GB     | 68.8 (92.6) | 82.1 (92.6)  | 75.4 (97.6) | 78.1 (95.0) | 63.4 (88.5) |
|                     | KNN    | 73.2 (73.0) | 76.7 (76.1)  | 92.7 (93.4) | 83.7 (83.7) | 55.4 (53.0) |
|                     | SVM    | 64.5 (74.9) | 84.6 (87.0)  | 63.6 (77.8) | 72.0 (82.0) | 65.7 (72.4) |

E. RESULTS ON LONG-BEACH-VA

In Fig. 5, PCA performs worst among the four feature selection methods and there is no apparent increment in accuracy. Different from the results in Hungarian, the GB classifier owns the best results in RF, Kendall, and mRMR. ANN, KNN, and RF have an apparent decline in accuracies. Besides, within these three selectors, the accuracies have a large fluctuation which means there exists strong interaction in the selected features. Generally, all the three selectors perform well on Long-Beach-Va, and mRMR performs better than other methods, its GB has the best accuracy, 99%.

F. SELECTED FEATURES

Features selected by Kendall, Random Forest, and mRMR algorithms on these three datasets are presented in Table 4.

We sort the features by their importance, the smaller serial number is more important in the above methods. To display the feature difference between the above algorithms and medical view, we highlight the processed features with bold in Kendall, Random Forest, and mRMR. The first 13 features provided by selectors are quite different from the medical view. For example, 3 (age), 4 (sex), 10 (resting blood pressure), and 12 (serum cholesterol in mg/dl) usually are behind 13. Therefore, we think that they may not play a vital role in the medical view. In contrast, 60 (ladprox), 61 (laddist), 63 (cxmain), and 67 (rcaprox) often appear in the first 13 columns in Kendall and Random Forest. According to the description of the dataset, these four features are recorded, but without explaining their meanings. They should contain some latent information that helps diagnose the disease.
FIGURE 2. Accuracy under the incremental heavy importance feature.

What’s more, in these first 13 features, mRMR contains the least number of features in the medical view, it’s also different from the Kendall and Random Forest methods.

G. COMPARISON

In Table 5, we assemble all results to have a direct comparison in datasets and feature selectors. Contrasting the results between LDA and 74 features without a selector, we found that LDA is no capacity in advancing prediction accuracy, it appears to be that LDA is not appropriate for the binary classification problem. PCA, Kendall, Random Forest, and mRMR can hugely reduce the number of selected features, and mRMR method should be the best one. Plus, the best methods of these feature selectors are the same on the three datasets. Concerning the best accuracies, different methods show their best performances on datasets. Compared with PCA’s results, the best classifier on Cleveland is ANN, accuracy increases 1.4%, feature number decreases 11; Hungarian is the Random Forest, the accuracy increases 4.9%, feature number increases 1; Long-Beach-Va is GB, the accuracy increases 11.7%, feature number decreases 17. We can also observe that the optimal feature number usually is below 17 after feature selection. What’s more, when we compare the results in the tables above together, nearly the first ten features can achieve more than 90% of the best performance value. With the same amount of features, the accuracy of mRMR is higher than other methods with the same classifier, which demonstrates that it can select features more accurately.
FIGURE 3. Results on Cleveland.

(a) Cleveland-PCA

(b) Cleveland-RF

(c) Cleveland-Kendall

(d) Cleveland-mRMR

FIGURE 4. Results on Hungarian.

(a) Hungarian-PCA

(b) Hungarian-RF

(c) Hungarian-Kendall

(d) Hungarian-mRMR
In Table 6, we compare our work with other researchers’ on the Cleveland datasets. Most of the researchers are prone to apply the dataset with 13 features. After the comparison, the mRMR not only has the best accuracy in these methods but also owns the least feature amount. Furthermore, the mRMR makes nearly all the classifiers achieve the highest accuracy.

**H. COMPLEXITY**

Assume the number of the instances and features are $n$ and $m$, respectively, and the selected feature number is $k$. The time complexity of PCA and LDA is $O(kn^2)$, Random Forest is $O(mn \log(n))$, SVM is $O(n^3)$, mRMR is $O(kmn)$, Kendall is $O(mn^2)$, KNN is $O(n)$, GB is $O(mn)$, and LR is $O(n^2)$. Relatively speaking, the number of total features $m$ and the selected features $k$ is small compared with instances, therefore, we pay main attention to the instance’s influence. For a clear view, we present several time complexity curves with the increment of instance in Fig. 6. It displays that the time complexity with the instances.
fourth line context: "These features may contain information that helps doctors diagnose disease. In the future, we plan to collect more large-scale datasets on heart disease and verify mRMR method on them.

**ACKNOWLEDGMENT**

The authors would like to thank the experts who have contributed in development of heart disease database.

**REFERENCES**

[1] R. O. Bowon, D. L. Mann, D. P. Zipes, and P. Libby, Braunwald’s Heart Disease e-Book: A Textbook of Cardiovascular Medicine. London, U.K.: Elsevier Health Sciences, 2011.

[2] J. Nahar, T. Imani, K. S. Tickle, and Y. P. P. Chen, “Association rule mining to detect factors which contribute to heart disease in males and females,” Expert Syst. Appl., vol. 40, no. 4, pp. 1086–1093, Mar. 2013.

[3] R. Mica, J. L. Peñalvo, F. Cudhea, F. Imamura, C. D. Rehm, and D. Mozaffarian, “Association between dietary factors and mortality from heart disease, stroke, and type 2 diabetes in the United States,” Jama, vol. 317, no. 9, pp. 912–924, 2017.

[4] E. J. Benjamin, P. Muntner, A. Alonso, M. S. Bittencourt, C. W. Callaway, A. P. Carson, A. M. Chamberlain, A. R. Chang, S. Cheng, S. R. Das, and F. N. Delling, “Heart disease and stroke statistics—2019 update: A report from the American heart association,” Circulation, vol. 139, no. 10, pp. e56–e528, 2019.

[5] B. Wu, “Hierarchical macro strategy model for MOBA game AI,” in Proc. AAAI Conf. Artif. Intell., 2019, vol. 33, no. 1, pp. 1206–1213.

[6] Y. Benkler, “Don’t let industry write the rules for AI,” Nature, vol. 569, no. 7754, pp. 161–162, 2019.

[7] S. E. Diliszian and E. L. Siegel, “Artificial intelligence in medicine and cardiac imaging: Harnessing big data and advanced computing to provide personalized medical diagnosis and treatment,” Current Cardiol. Rep., vol. 16, no. 1, p. 441, Jan. 2014.

[8] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, “Artificial intelligence in healthcare: Past, present and future,” Stroke Vascular Neurol., vol. 2, no. 4, pp. 1–14, Dec. 2017.

[9] A. Esteve, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, M. H. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” Nature, vol. 542, pp. 115–118, Feb. 2017.

[10] R. Atallah and A. Al-Moussa, “Heart disease detection using machine learning majority voting ensemble method,” in Proc. 2nd Int. Conf. New Trends Comput. Sci. (ICTCS), Oct. 2019, pp. 1–6.

[11] F. Tan, X. Fu, Y. Zhang, and A. G. Bourgeois, “A genetic algorithm-based method for feature subset selection,” Soft Comput., vol. 12, no. 2, pp. 111–120, Sep. 2007.

[12] V. Singh, N. K. Verma, and Y. Cui, “Type-2 fuzzy PCCA approach in extracting salient features for molecular cancer diagnostics and prognostics,” IEEE Trans. Nanobiosci., vol. 18, no. 3, pp. 482–489, Jul. 2019.

[13] T. Karayilan and Ö. Kılıç, “Prediction of heart disease using neural network,” in Proc. Int. Conf. Comput. Sci. Eng. (UBMK), Oct. 2017, pp. 719–723.

[14] B. Kolukisa, H. Hacilar, G. Goy, M. Kas, B. Bakır-Gungor, A. Aral, and V. C. Gungor, “Evaluation of classification algorithms, linear discriminant analysis and a new hybrid feature selection methodology for the diagnosis of coronary artery disease,” in Proc. IEEE Int. Conf. Big Data (Big Data), 2018, pp. 2232–2238.

[15] H. Shafizadeh-Moghadam, “Fully component selection: An efficient combination of feature selection and principal component analysis to increase model performance,” Expert Syst. Appl., vol. 186, Dec. 2021, Art. no. 115678.

[16] A. Y. Ng, “Feature selection, L1 vs. L2 regularization, and rotational invariance,” in Proc. 21st Int. Conf. Mach. Learn. (ICML), 2004, p. 78.

[17] G. Chandrashekar and F. Sahin, “A survey on feature selection methods,” Comput. Elect. Eng., vol. 40, no. 1, pp. 16–28, Jan. 2014.

[18] I. K. A. Enriko, M. Suryanegara, and D. Gunawan, “Heart disease prediction system using k-Nearest neighbor algorithm with simplified patient’s health parameters,” J. Telecommun., Electron. Comput. Eng., vol. 8, no. 12, pp. 59–65, 2016.
[19] P. E. N. Lutu and A. P. Engelbrecht, “A decision rule-based method for feature selection in predictive data mining,” Expert Syst. Appl., vol. 37, no. 1, pp. 602–609, Jan. 2010.

[20] W. You, Z. Yang, and G. Ji, “Feature selection for high-dimensional multi-category data using PLS-based local recursive feature elimination,” Expert Syst. Appl., vol. 41, no. 4, pp. 1463–1475, Mar. 2014.

[21] J. S. Deogun, S. K. Choubey, V. V. Raghavan, and H. Sever, “Feature selection and effective classifiers,” J. Amer. Soc. Inf. Sci., vol. 49, no. 5, pp. 423–434, 1998.

[22] F. Amini and G. Hu, “A two-layer feature selection method using genetic algorithm and elastic net,” Expert Syst. Appl., vol. 166, Mar. 2021, Art. no. 114072.

[23] C. De Stefano, F. Fontanella, and C. Marrocco, “A GA-based feature selection algorithm for remote sensing images,” in Proc. Workshops Appl. Evol. Comput. Cham, Switzerland: Springer, 2008, pp. 285–294.

[24] C. E. Shannon, “A mathematical theory of communication,” Bell Syst. Tech. J., vol. 27, no. 3, pp. 379–423, Jul./Oct. 1948.

[25] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[26] X. Yan and M. Jia, “Intelligent fault diagnosis of rotating machinery using improved multiscale dispersion entropy and mRMR feature selection,” Knowl.-Based Syst., vol. 163, pp. 450–471, Jan. 2019.

[27] F. Özyurt, “A fused CNN model for WBC detection with MRMR feature selection and extreme learning machine,” Soft Comput., vol. 24, pp. 8163–8172, Jun. 2019.

[28] Y. Eroglu, M. Yildirim, and A. Cinar, “mRMR-based hybrid convolutional neural network model for classification of Alzheimer’s disease on brain magnetic resonance images,” Int. J. Imag. Syst. Technol., vol. 32, no. 2, pp. 517–527, Mar. 2022.

[29] N. Kumar and D. Kumar, “IMGWO based ANN: A new heart disease diagnosis model to classify real world dataset,” Tech. Rep. [Online]. Available: http://www.ijcse.com/abstract.html?file=21-12-04-182, doi: 10.21817/ijcse/2021/v12i4/211204182.

[30] P. M. Domingos, “A few useful things to know about machine learning,” Commun. ACM, vol. 55, no. 10, p. 78–87, Oct. 2012.

[31] A. M. Martínez and A. C. Kak, “PCA versus LDA,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 2, pp. 228–233, Feb. 2001.

[32] D. K. Choubey, M. Kumar, V. Shukla, S. Tripathi, and V. K. Dhandhania, “Comparative analysis of classification methods with PCA and LDA for diabetes,” Current Diabetes Res., vol. 16, no. 8, pp. 833–850, Sep. 2020.

[33] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[34] D. Dua and C. Graff, “UCI machine learning repository,” School Inf. Comput. Sci., Univ. California, Irvine, CA, USA, 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[35] D. Khanna, R. Sahu, V. Baths, and B. Deshpande, “Comparative study of classification techniques (SVM, logistic regression and neural networks) to predict the prevalence of heart disease,” Int. J. Mach. Learn. Comput., vol. 5, no. 5, p. 414, 2015.

[36] S. S. Khan and S. Quadri, “Prediction of angiographic disease status using rule based data mining techniques,” Biol. Forum, Int. J., vol. 8, no. 2, pp. 103–107, 2016.

[37] D. R. V. S. Kodati, “Analysis of heart disease using in data mining,” Expert Syst. Appl., vol. 41, no. 4, pp. 1463–1475, Mar. 2014.

[38] K. Uyar and A. Ilhan, “Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks,” Proc. Comput. Sci., vol. 120, pp. 588–593, Jan. 2017.

[39] A. Gupta, R. Kumar, H. S. Arora, and B. Raman, “MIFH: A machine intelligence framework for heart disease diagnosis,” IEEE Access, vol. 8, pp. 14659–14674, 2020.

**GAOSHUAI WANG** is currently pursuing the Ph.D. degree with UTBM. His research interests include machine learning, computer vision, optimization, and heuristic learning.

**FABRICE LAURI** received the Ph.D. degree in computer science from the Université de Nancy 2, in 2004.

He has been an Associate Professor with the Université de Technologie de Belfort-Montbéliard (UTBM), since September 2006. He leads his research at the “Institut de Recherche sur les Transports, l’Énergie, et la Société” (IRTES-SeT), Belfort, France. His research interests include optimization problems in multi-agent systems, that is systems involving several autonomous entities, metaheuristics inspired by nature, like ant colony algorithms or evolutionary algorithms, and machine learning algorithms, especially those from reinforcement learning, are part of his favorite techniques for solving such problems.

**AMIR HAJJAM EL HASSANI** received the Ph.D. degree in computer science from the Université de Haute Alsace (UHA), in 1990.

He is an Associate Professor HDR with the Université de Technologie de Belfort-Montbéliard (UTBM) and the Head of software engineering. He is the Deputy Director of Nanomedicine Laboratory, Imaging & Therapeutics of Université de Franche Comté (UFC), and the Research Team Leader of Health Systems Organization. His research interests include data mining and machine learning for decision support in the field of e-Health. He is a member of the Science Steering Committee of the annual conferences IADIS e-Health and e-Medsys. He was the Co-Chair of the First International Conference eTelemed’09 and has been its Advisory Chair, since 2010. He has organized many conferences and chaired several technical sessions. He is an Expert to the ANRT France, AIRI Alsace, and CNRST Morocco. He is an editorial board member of four international journals and the author/coauthor of three books and many international publications in refereed journals and conferences.

---

**G. Wang et al.: Feature Selection by mRMR Method for Heart Disease Diagnosis**