Differential Evolution Based Feature Selection: A Niching-based Multi-objective Approach

Peng Wang, *Graduate Student Member, IEEE*, Bing Xue, *Senior Member, IEEE*, Jing Liang, *Senior Member, IEEE*, and Mengjie Zhang, *Fellow, IEEE*

***Abstract*—Feature selection is to reduce both the dimensionality of data and the classification error rate (i.e., increase the classification accuracy) of a learning algorithm. The two objectives are often conflicting, thus a multi-objective feature selection method can obtain a set of non-dominated feature subsets where each has a different size and a corresponding classification error rate. However, most existing feature selection algorithms have ignored that, for a given size, there can be different feature subsets with very similar or the same accuracy. This paper introduces a niching-based multi-objective feature selection method that simultaneously minimizes the number of selected features and the classification error rate. The proposed method conceives to identify: 1) a set of feature subsets with good convergence and distribution, 2) multiple feature subsets that maximize the classification accuracy of a given learning algorithm, i.e., choose the same number of features with almost the same lowest classification error rate. The contributions of this paper are threefold. Firstly, a niching and global interaction mutation operator is proposed that can produce promising feature subsets. Secondly, a newly developed environmental selection mechanism that relaxes the Pareto-dominance relationship can store the equal informative feature subsets. Lastly, the proposed subset repairing mechanism can generate better feature subsets and further remove the redundant features. The proposed method is compared against six multi-objective feature selection algorithms on 19 datasets including both binary and multi-class classification tasks. The results show that the proposed method can evolve a rich and diverse set of non-dominated solutions for different feature selection tasks, and their availability helps in understanding the relationships between features.**

***Index Terms*—Differential evolution (DE), feature selection, classification, evolutionary multi-objective optimization.**

I. INTRODUCTION

Classification is one of the main tasks of machine learning. A classification task typically involves training a learning algorithm using labelled training instances. The trained classifier is then used to predict the class labels of previously unseen instances (i.e., test instances). The training phase of the classification algorithm aims to discover the relationship between instances' properties (i.e., features) and their class labels. Therefore, the classification performance of the learned algorithm is directly affected by the quantity and the quality of features [1]. The existence of redundant features, irrelevant features, and noisy features obscures the useful information from relevant or discriminating features and thereby degrades the performance of classification. By removing irrelevant and redundant features, feature selection can reduce the number of features and improve the quality of features [2].

Although being studied for decades, feature selection remains a challenging task because of the huge search space especially on high-dimensional data and the complex interactions among features [3]. A dataset with n original features can have 2^n possible feature subsets, i.e., the size of search space increases exponentially as the number of features. Furthermore, due to the complex interactions among features, different feature subsets could have similar or equal discriminating ability [4], [5]. Features can be divided into three categories: strongly relevant, irrelevant, and weakly relevant features, based on their relevance [6]. Strong relevance means a feature cannot be deleted from the feature subset without affecting the classification performance. Irrelevance indicates removing a feature will not affect the classification performance at all. Weak relevance indicates that a feature is not always necessary but may become useful when the feature is combined with other features, although each of them may contain only a little information about the class labels. An optimal feature subset often includes strongly relevant and weakly relevant but non-redundant features [7]. Different divisions of strongly and/or weakly relevant features, e.g., redundant and non-redundant features, could be generated which results in different feature subsets with very similar or the same classification performance.

The existence of different solutions with almost the same classification performance [8]–[12] is called many-to-one problems in [13], identifying equally informative subsets in [4], and multimodal optimization problems in [14]–[17]. Most existing feature selection methods do not consider or keep the candidate solutions with similar or the same classification performance during the learning process, and thus optimal feature subsets can be lost. Furthermore, it is essential to understand different feature subsets with the same classification performance. In disease diagnosis and biomarker detection, Liu et al. [9] observed that two different feature subsets in the DLBCL dataset [18], $\{$*X62078_at*, *L33842_rna1_at*, *J02645_at*$\}$ and $\{$*X56494_at*, *M57710_at*, *U19495_s_at*$\}$, can achieve the same classification accuracy (91.7%). The gene (i.e., feature), *L33842_rna1_at*, in the first subset, is over-expressed in the group of the DLBCL patients, while *M57710_at* in the second subset is a significantly upregulated gene in the DLBCL lymphoma samples [9]. Due to the existence of both feature subsets with the same classification performance, two distinct functional modules are likely to distinguish normal individuals

M. Zhang, B. Xue and P. Wang are with the Evolutionary Computation Research Group, Victoria University of Wellington, Wellington 6140, New Zealand.

J. Liang is with the School of Electrical Engineering, Zhengzhou University, Zhengzhou 450001, China. (*Corresponding author:Jing Liang*)

from DLBCL patients.

Classical feature selection methods typically can only output one feature subset, although most of them consider the presence of redundant and irrelevant features. Some methods such as ReliefF [19] require a user to have domain knowledge, such as providing the number of selected features in advance. Moreover, combining several highly ranked features may produce redundancy, which may result in the method not achieving the desired classification accuracy. Evolutionary computation (EC) methods especially evolutionary multi-objective optimization (EMO) methods can overcome those limitations because EC methods require no domain knowledge and assumptions about the search space [2]. Meanwhile, EMO methods can find a set of trade-off (between the subset size and the classification error rate) solutions in a single run using their population-based search mechanism [20]. Furthermore, EMO with niching techniques has been the leading approach in searching for multiple optimal solutions of many real-world multimodal problems including feature selection [10], [15], [21], [22].

For a niching-based EMO method, one of the important aspects is the search algorithm which aims to produce promising individuals. Many different methods have been proposed and commonly used in the literature, e.g., genetic algorithms (GAs) [23], particle swarm optimization (PSO) [24], and differential evolution (DE) [25]. Among them, DE is used by many researchers because of its simplicity and effectiveness [17]. More importantly, DE variants have shown their superiority for solving multimodal optimization problems [13], [17], [26].

Existing niching-based feature selection methods also have limitations. Firstly, to produce promising candidate feature subsets while avoid getting trapped into local optima, the quality of the parental individuals from both the niche and the whole population should be considered [17], [26]. However, many existing methods ignore this point. Secondly, to find and maintain multiple equivalent feature subsets, applying crowding estimations or density measures as an environmental selection scheme in both the search space and the objective space is needed [13], [16]. However, most studies do not consider the intrinsic characteristics of feature selection itself. For example, the dimensionality of the search space or the solution space of a feature selection task is significantly larger than that of the objective space, which is different from many existing multimodal multi-objective benchmark function optimization problems in [16]. Therefore, many existing methods can hardly find and maintain multiple equivalent solutions especially for a feature selection task with thousands or even tens of thousands of features. Lastly, even if a few different feature subsets with similar or the same classification performance can be found arbitrarily, the obtained solutions may have redundant features. For example, [5] shows that two different feature subsets $\{feathers, milk, airborne, toothed\}$ and $\{milk, airborne, toothed, Backbone\}$ with KNN in the Zoo dataset from UCI Machine Learning Repository [18] are observed to have the same classification accuracy of 90.0%. However, when combining the common features from both feature subsets into a new feature subset, i.e., subset $\{milk, airborne, toothed\}$, using $\{milk, airborne, toothed\}$ can also reach the classification accuracy of 90.0%. This indicates that the original two feature subsets obtained still include some redundant features, e.g., $feathers$ or $Backbone$. Those motivate us to design a new operator to further remove the redundant features while retaining the non-redundant features during the evolutionary learning process. Moreover, some redundant features, if combined with complementary features, could allow further improvements of a niching-based EMO feature selection method. Therefore, this work aims to propose a niching-based multi-objective DE method for feature selection (termed NMDE) to addresses the above limitations. The major contributions are as follows:

- Under the framework of EMO, a new mutation operator combining the local information of the niche (the neighbors of an individual) and the global information of the whole population is proposed to locate the regions with good feature subsets and maintain global search ability. The results reveal that the proposed mutation operator can not only speed up the convergence but also produce good feature subsets with lower classification error rate.
- NMDE transforms the task of searching for multiple optimal feature subsets into searching for ψ-quasi equal feature subsets to handle the issue that multiple feature subsets with the same size may have very close but different classification accuracy. The term, ψ-quasi equal, defines the closeness of the classification accuracy between two feature subsets. NMDE also relaxes the traditional Pareto-dominance concept and improves the calculations of the crowding estimation involving both the solution space and the objective space. The results show that the proposed environmental selection scheme can help NMDE find and maintain ψ-quasi equal feature subsets.
- The possible situations of different feature subsets with similar or the same classification performance (i.e., ψ-quasi equal feature subsets) are divided into three situations. The proposed subset repairing mechanism considering all these three situations can produce better feature subsets over the ψ-quasi equal feature subsets and remove redundant features in the ψ-quasi equal feature subsets. The results indicate that the proposed subset repairing mechanism can improve the classification performance of the algorithm and decrease the redundant rate of the obtained feature subsets.

The remainder of this paper is structured as follows. Section II describes the related work on feature selection. Section III presents the details of the proposed NMDE algorithm. Section IV shows the experimental settings. The results are then discussed in Section V. Finally, Section VI concludes this paper.

II. Related Work

Researchers have developed many feature selection methods for classification tasks. Some conventional feature selection methods are introduced in this section. Additionally, this section reviews some EC-based multi-objective feature selection methods, and the main focus is on the EMO-based feature selection methods.

A. Conventional Feature Selection Methods

Feature selection methods can be roughly classified into wrapper, filter, and embedded methods. Wrapper methods evaluate the goodness of a feature subset utilizing a specific learning algorithm (i.e., classifier), while filter methods rely on some statistical measures of the training data without using any learning algorithm. In embedded methods, feature selection is embedded into the learning process of a classifier. In comparison with a filter method, a wrapper method is said to be able to achieve higher classification accuracy [3]. Therefore, a wrapper approach is employed in this work to evaluate the candidate feature subsets.

The ReliefF [27] and FOCUS [28] methods are two classical filter feature selection algorithms. FOCUS exhaustively examines all possible feature subsets and then chooses the one that has the smallest number of features with the best classification performance. Due to the exhaustive search nature, FOCUS has a high computational cost. To overcome this limitation, correlated feature selection (CFS) was proposed in [29]. To deal with the problem with high-dimensional data, a fast correlation-based filter method (FCBF) was proposed in [30]. The limitations of ReliefF explained in the fourth paragraph of Section I.

Two typical examples of wrapper feature selection methods are sequential forward selection (SFS) and sequential backward selection (SBS) [31]. However, both methods undergo the so-called *nesting effect* because a feature that is selected (or eliminated) cannot be eliminated (or selected) later. To overcome this limitation, Pudil et al. [32] proposed sequential backward floating selection (SBFS) and sequential forward floating selection (SFFS). In addition, Peng et al. [33] proposed the minimum Redundancy Maximum Relevance (mRMR) principle to select features. The key idea is to select the features with the highest relevance to the target class while the lowest redundancy among the features. However, SBFS, SFFS, and mRMR may fall into a local optimum.

Recently, the sparse learning-based feature selection methods have attracted a lot of attention [34], [35], such as the sparse multinomial logistic regression model (SBMLR) [36] and the robust feature selection method (RFS) [37]. These methods try to find the optimal weight vector, which minimizes the measurement error along with some regularization terms. Due to the sparse regularizer, some learned weights can be quite small, and their corresponding features will be discarded. However, these methods usually require the number of selected features in advance.

B. EC-based Feature Selection Methods

An EMO method for solving feature selection problems aims to obtain a widely distributed Pareto front (PF) of non-dominated solutions (i.e., feature subsets), as shown in Fig. 1 in the objective space, to meet different requirements of decision-makers. Moreover, the algorithm is usually required to converge fast due to the limited budget on the fitness evaluations (FEs). To better review the related works on these efforts to solve feature selection problems, the following aspects are described.

1) Environmental Selection Strategies: The term, *environmental selection*, is usually used in EMO methods, which is to select individuals to form the population into the next generation. The concept of Pareto-dominance and crowding estimation [38] are the most widely used environmental selection strategies in EMO methods including multi-objective feature selection methods [39]–[43]. However, these algorithms may obtain many feature subsets around centre of PF [44], [45]. To cover this issue, different dominance relationships (e.g., grid-dominance [46], ε-dominance [47], and fuzzy dominance [48]) have been applied to feature selection, and some improved calculations of crowding distance are also proposed [10], [13]. Although the relaxed forms of the Pareto-dominance, i.e., the grid-dominance and the ε-dominance, can regulate the granularity of the approximation of objective values of feature subsets, a large number of good feature subsets may be lost, if the ε value or the number of grids is badly chosen. For the crowding estimation, Yue et al. [26] considered the solutions that are ranked higher than the current solution in non-dominated sorting when calculating the crowding distance of a solution rather than considering each front separately (e.g., NSGA-II), and proposed the MMODE_ICD algorithm. Although MMODE_ICD obtained promising results on addressing multimodal multi-objective benchmark optimization problems, some parameters, e.g., the selection ratio of solutions from different fronts, are not easy to set when applying MMODE_ICD to feature selection. Hu et al. [48] proposed a fuzzy dominance relationship and a fuzzy crowding distance measure with multi-objective PSO for feature selection.

Another issue in the environmental selection process is the duplication of candidate feature subsets (e.g., S_2 and S_3 in Fig. 1), which will decrease the population diversity and cause premature convergence. In [49], a features selection task is treated as a sparse multi-objective optimization problem. The proposed SparseEA algorithm found the duplicated feature subsets, but it did not design a specific strategy to modify them. Xu et al. [50] designed an offspring modification mechanism to improve the population diversity by using the features' selected frequency of feature subsets sitting at the first front in the current generation. However, the quality of the produced feature subsets cannot be guaranteed. Wang et al. [46] proposed a subset filtering mechanism to choose a feature subset from the duplicated ones according to their confidence scores. Xu et al. [20] addressed the issues of the *decision vector duplication* and the *objective vector duplication*, and proposed a duplication analysis based feature selection method. However, the above methods ignore the existence of multiple optimal feature subsets in a feature selection task.

2) Finding Multiple Optimal Feature Subsets: The existence of multiple optimal feature subsets indicates that picking one of them may not be a good choice, and the users are likely to make a choice according to their domain knowledge [9]. To search for multiple optimal feature subsets, Wang et al. [5] employed EMO technique to address feature selection tasks. The key idea is to assign an equally important crowding metric to different feature subsets with the same classification performance while assigning 0 to the duplicated solutions. The results showed that the proposed method can produce different

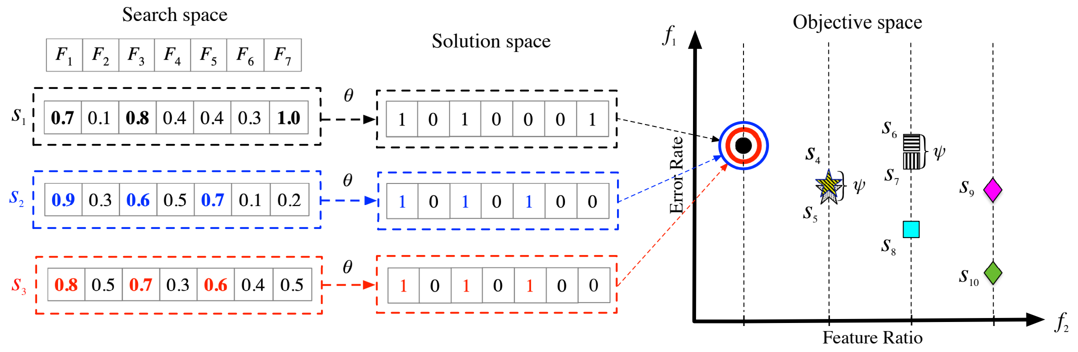

Fig. 1. The situations of ψ-quasi equal feature subsets in the search space, the solution space, and the objective space. Suppose a dataset has seven features, F_1 to F_7, and S_1 to S_{10} represent ten feature subsets (i.e., solutions) on this feature selection task. According to an individual's representation in Section III.A, a solution in the search space is coding with real numbers. By comparing each element with θ, a selected feature will be indicated by 1 otherwise 0 in the solution space. In the objective space, S_1 to S_3 are sitting at the same point since they have the same objective values. Based on the definition in Section III.B, S_4 is ψ-quasi equal subset to S_5, and S_6 is ψ-quasi equal subset to S_7.

feature subsets with the same classification performance but without significantly improving the performance regarding hypervolume (HV) and inverted generational distance (IGD). Karakaya et al. [4] proposed two feature selection methods, a wrapper for quasi equally informative subset selection (W-QEISS) and a filter for quasi equally informative subset selection (F-QEISS), which use EMO techniques with extreme learning machine to find multiple optimal feature subsets. However, the study in [4] lacks comparative experiments with other PSO-based or DE-based EMO methods.

A main approach in searching for multiple solutions with the same objective value(s) is to use niching techniques [15]. Niching techniques, including crowding-based [51], sharing-based [52], and speciation-based [53] methods, are to incorporate the neighboring information of the evolved population into the search process. Kamyab et al. [11] proposed a dynamic fitness sharing based PSO method for feature selection. However, the performance of the proposed method is sensitive to the niche count. Yue et al. [10] applied the ring topology and a multi-objective PSO algorithm to address feature selection tasks, namely MO_Ring_PSO_SCD. The proposed special crowding distance (SCD) mechanism considers the crowding estimations both in the search space and the objective space. Kanchan et al. [54] further improved the method in [10] and developed a filter-based multimodal feature selection method. However, both methods from [10], [54] need a large population to introduce niche.

Although several niching-based feature selection methods have been proposed, the feature subsets obtained from these methods such as in [10] and [54] still include a number of redundant features. In the following section, we proposed a new method named NMDE, which employs both EMO and niching techniques to search for multiple optimal feature subsets and further remove possible redundant features.

III. Proposed Method

This section introduces the proposed niching based multi-objective DE (NMDE) feature selection method, including the representation, the objective functions, the equivalence of feature subsets, and algorithmic components of NMDE in detail.

A. Representation and Objective Functions

The representation of an individual (shown in Fig. 1) in this work is a vector with real value encoding (between 0 and 1). The dimensionality of each vector is equal to the number of original features in the dataset, where a pre-defined threshold θ is employed to determine whether one feature is chosen. If the corresponding value in the vector (i.e., position entry) is not less than the threshold θ, the feature will be selected, otherwise not be selected.

The objective functions are shown in Eq. (1), which are to minimize the classification error rate (ER) and the number of selected features simultaneously.

$$\min \begin{cases} f_1 = ER = \frac{FP+FN}{TP+TN+FP+FN} \\ f_2 = FR = \frac{\#Selected\ Features}{\#\text{Original Features}} \end{cases} \tag{1}$$

where FP, FN, TP, and TN are the false positives, false negatives, true positives, and true negatives, respectively. ER is the error rate and FR is the ratio of the number of selected features (i.e., $\#Selected\ Features$) over the total number of original features (i.e., $\#\text{Original Features}$).

B. Equivalence of Feature Subsets

Definition 1: For a given measure of the classification error rate (i.e., f_1), the ratio of the number of features selected in a feature subset (i.e., f_2), and a set of feature subsets $\mathbb{S}$, two feature subset S_1 and S_2 from $\mathbb{S}$ are considered having similar or the same classification performance with respect to a given learning algorithm, i.e., feature subset S_1 is a ψ-quasi equal feature subset to feature subset S_2, if $f_2(S_1) = f_2(S_2)$ and $|f_1(S_1) - f_1(S_2)| \leq \psi$, ψ is a small constant, $0 \leq \psi < 1$. Noted that $\psi = 0$ means that different feature subsets have exactly the same f_1 and f_2 values, e.g., S_1, S_2, S_3 in Fig. 1.

As shown in Fig. 1, feature subsets from the following three groups, $\{S_1, S_2, S_3\}$, $\{S_4, S_5\}$, and $\{S_6, S_7\}$, are ψ-quasi equal feature subsets. One note is that feature subset S_8 has a lower classification error rate than both S_6 and S_7, thus S_8 is preferred in NMDE during the evolutionary training process.

In the following, the proposed NMDE algorithm considering ψ-quasi equal feature subsets will be introduced in detail.

C. Niching Behavior

In the proposed NMDE method, each individual $\vec{x}_i$ has its own niche. Specifically, $\vec{x}_i$ combined with its N nearest surrounding neighbors form a niche (local area), termed $\mathcal{N}_i$, $|\mathcal{N}_i| = N$. Hamming distance is used to calculate the distance between paired individuals in the solution space. N can be set to 8, which is motivated by the promising results from [17]. Along this way, the developed mutation operator enables the current individual $\vec{x}_i$ to learn knowledge from the niche set $\mathcal{N}_i$ or the whole population P.

D. Mutation

Classic mutation operators in DE, e.g., DE/$rand$/1 and DE/$rand$/2, have some limitations [17], since most of them randomly choose individuals from the whole population to generate new individuals, without concerning the quality or the distance of the selected individuals. Ideally, the search space should be fully explored to get the optimal solutions. A common way is to use niching information, i.e., the differential vector in mutation can be generated by individuals from the same niche, but it may also increase the risk of getting trapped into local optima.

To speed up the convergence of the population and avoid getting trapped into local optima, the developed mutation operator in Eq. (2) combines the local information of the niche ($\mathcal{N}_i$) and the global information of the whole population (P).

$$\vec{v}_i = \begin{cases} \vec{x}_i + F \times (\vec{x}_{r_1} - \vec{x}_{r_2}) & \\ \quad r_1 \neq r_2 \neq i \in \mathcal{N}_i, & \text{if } \vec{x}_i \in \mathcal{S}_i \\ \vec{x}_i + F \times (\vec{x}_{nal} - \vec{x}_i) + F \times (\vec{x}_{g_{r_1}} - \vec{x}_{g_{r_2}}) & \\ \quad g_{r_1} \neq g_{r_2} \neq i \neq nal \in P, & \text{otherwise} \end{cases} \tag{2}$$

where F is the mutation factor, and r_1 and r_2 are randomly selected from the niche $\mathcal{N}_i$ of $\vec{x}_i$. Meanwhile, g_{r_1} and g_{r_2} are randomly selected from the whole population P. $\mathcal{S}_i$ stores the individuals in the first front in $\mathcal{N}_i$, and $\vec{x}_{nal}$ is the individual in $\mathcal{S}_i$ with the lowest classification error rate.

Two different situations are considered in the proposed mutation operator. One situation is $\vec{x}_i \in \mathcal{S}_i$ which means that individual $\vec{x}_i$ is sitting at the first front in its niche in the current generation. In this case, $\vec{x}_i$ might be close to an optimum. Therefore, two individuals from the same niche $\mathcal{N}_i$ are randomly selected to generate local exploitation information for $\vec{x}_i$. The other situation is $\vec{x}_i \notin \mathcal{S}_i$, meaning that $\vec{x}_i$ is dominated by the individual(s) in the niche. In this case, $\vec{x}_{nal}$ is used to guide $\vec{x}_i$, so that the individual $\vec{x}_i$ can find the potential optima quickly. Meanwhile, to prevent individuals from being trapped into local optima, the global information is also used, i.e., g_{r_1} and g_{r_2} are randomly selected from the whole population P.

E. Subset Repairing Scheme

The proposed subset repairing scheme aims to produce better feature subsets based on the ψ-quasi equal feature subsets to improve the classification performance. Noted that two or more duplicated feature subsets (e.g., S_1 and S_3 in Fig. 1) can achieve the same objective values, thus the duplicated feature subsets should be removed first. After that, the presence of the ψ-quasi equal feature subsets can be divided into three situations. Fig. 1 shows *Situation* 1 where some common features are included in the ψ-quasi equal feature subsets (i.e., both features F_1 and F_3 are included in S_1 and S_2). It is important to check the classification performance of the common features included since it may help the algorithm further remove the possible redundant feature(s) from the ψ-quasi equal feature subsets. *Situation* 2 means that no common feature is included among the ψ-quasi equal feature subsets, as shown in the left side of Fig. 2. Several different feature combinations are ψ-quasi equal feature subsets. For *Situation* 2, combining all features included in different feature subsets as a new feature subset may produce a lower classification error rate. The most complicated one (*Situation* 3) is shown in the right side of Fig. 2 where some solutions has common features.

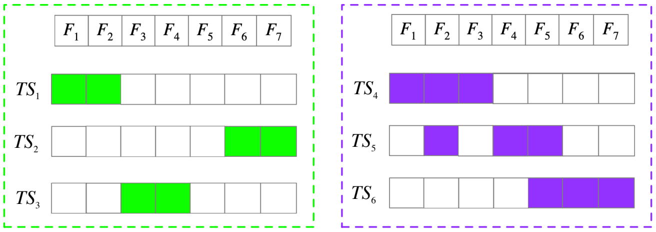

Fig. 2. Another two situations of ψ-quasi equal feature subsets.

Algorithm 1: Subset Repairing Operator

```
   Input: $P$, $O$, $P_u$, and $\psi$
   Output: $PO$,$P_u$, and $Qe$
1  begin
2      Get $PO = P \bigcup O$ and $PO\_fit$,
3      Remove the duplicated solutions in $PO$,
4      Get $Qe$ (storing multiple groups of the $\psi$-quasi equal feature
         subsets) via Alg. 2,
5      for $i$ = 1, . . . , $|Qe|$ do
6          Get $t$ equally feature subsets, $TS_1$-$TS_t$, in $Qe(i)$,
7          Calculate the selected frequency of features, termed $freq$,
8          if $\max\{freq_j\} > 1/t$ then
9              Produce a new solution $Sf$ via Eq. (3),
10         else
11             Produce a new solution $Sf$ via Eq. (4),
12         end
13         if $Sf \notin P_u$(unique set) then
14             $PO = PO \bigcup Sf$ and $P_u = P_u \bigcup Sf$,
15         end
16     end
17 end
```

Considering the above three situations, the proposed subset repairing operator is presented in Alg. 1. After producing new individuals, a set (termed PO) combining parents (P) and offspring (O) is formed (Line 2 in Alg. 1). The duplicated feature subsets detected in the solution space from PO are firstly removed (Line 3). Then, a set Qe storing multiple groups of the ψ-quasi equal feature subsets is the output of Alg. 2. Suppose that the number of ψ-quasi equal feature subsets in $Qe(i)$ is t, termed $TS_1, TS_2, \ldots, TS_t$ (Line 6 in Alg. 1). Next, the selected frequency of each feature is computed based on $\{TS_1, TS_2, \ldots, TS_t\}$, termed $freq$ (Line 7 in Alg. 1).

For *Situations* 1 and 3 (Line 8 in Alg. 1), a new feature

Algorithm 2: ψ-quasi Equal Subsets Grouping

```
    Input: $PO$, $PO\_fit$, and $\psi$
    Output: $Qe$
1   begin
2       Divide $PO$ into $m$ groups ($Sz_1$, $Sz_2$, ..., $Sz_m$) via feature
          subsets' $FR$ (i.e., $f_2$) values,
3       Construct an empty set $Qe$ with length $m$,
4       for $i = 1, \ldots, m$ do
5           Get the feature subset $S_{\min}$ with the lowest $ER$ (i.e., $f_1$)
              value in $Sz_i$,
6           if $TS_1, TS_2, \ldots, TS_t$ in $Sz_i$ are $\psi$-quasi equal subsets to
              $S_{\min}$ then
7               $Qe(i) = TS_1 \bigcup TS_2 \bigcup \cdots \bigcup TS_t$
8           end
9       end
10      Remove the empty subset in $Qe$,
11  end
```

subset Sf is created by setting its j-th variable Sf_j with the following formula:

$$Sf_j = \begin{cases} \text{rand}[\theta, 1], & \text{if } freq_j > 1/t \\ \text{rand}[0, \theta), & \text{otherwise} \end{cases} \tag{3}$$

where $\text{rand}[l, u]$ means a randomly generated number between l and u. Meanwhile, $freq_j$ is the value of the j-th position in $freq$, and t is the number of ψ-quasi equal feature subsets.

Eq. (3) shows that one feature will be selected from Sf only when the selected frequency value of the feature is larger than $1/t$. Therefore, Sf will select fewer number of features than each feature subset in $\{TS_1, TS_2, \ldots, TS_t\}$. If the classification error rate of Sf is not larger than that of a solution in $\{TS_1, TS_2, \ldots, TS_t\}$, that means redundant features are included in $\{TS_1, TS_2, \ldots, TS_t\}$. During the environmental selection process, Sf will be preferred. Otherwise, all the solutions in $\{TS_1, TS_2, \ldots, TS_t\}$ are preferred.

For *Situation* 2 that means $\max\{freq_j\} \leq 1/t$ in Line 10 of Alg. 1, Sf will be created by combining all the features selected in $\{TS_1, TS_2, \ldots, TS_t\}$.

$$Sf_j = \begin{cases} \text{rand}[\theta, 1], & \text{if } freq_j > 0 \\ \text{rand}[0, \theta), & \text{otherwise} \end{cases} \tag{4}$$

Eq. (4) shows that one feature will be selected from Sf once the selected frequency value of the feature is higher than 0. Although the number of features selected in Sf is larger than that of a solution in $\{TS_1, TS_2, \ldots, TS_t\}$, Sf may have a lower classification error rate since the combined features may provide different complementary information towards the class.

To save the computational resources, all feature subsets produced and their objective values will be preserved as a set (termed P_u). If Sf is included in P_u, Sf will not enter into PO. Otherwise, the objective values of Sf will be calculated, and Sf is required to enter into both P_u and PO (Line 14 in Alg. 1).

F. Environmental Selection

In NMDE, the environmental selection shown in Alg. 3 is customized to keep the ψ-quasi equal feature subsets. The key idea is to relax the Pareto-dominance relationship to drag the ψ-quasi equal feature subsets into the same front.

Algorithm 3: Environmental Selection

```
    Input: $PO$, $Qe$, and $\psi$
    Output: $P$
1   begin
2       Rank $PO$ into $k$ fronts ($\mathcal{W}_1$, $\mathcal{W}_2$, ..., $\mathcal{W}_k$) via non-dominated
          sorting, and set $P = \varnothing$,
3       Update the non-dominated sorting via Alg. 4,
4       Get the crowding distances of the feature subsets in $\mathcal{W}_m$ via
          Eq. (6), $m \in [1, k]$,
5       Select $NP$ solutions to $P$ from $\mathcal{W}_1$-$\mathcal{W}_k$,
6   end
```

Algorithm 4: Update Sorting

```
    Input: $\mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_k$ and $Qe$
    Output: $\mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_k$
1   begin
2       for $i = 1, \ldots, |Qe|$ do
3           Get $TS_1, TS_2, \ldots, TS_t$ in $Qe(i)$,
4           Get the lowest front number of non-dominated sorting of
              the $t$ solutions, termed $r_{\min}$,
5           Calculate $ln$ via Eq. (5),
6           if $t > ln$ then
7               Calculate the crowding distances of the $t$ feature
                  subsets via Eq. (8),
8               Sort the $t$ feature subsets in a descending order,
9               Put the top $ln$ feature subsets to $\mathcal{W}_{r_{\min}}$,
10          else
11              Put all $t$ feature subsets to $\mathcal{W}_{r_{\min}}$,
12          end
13      end
14  end
```

Firstly, all individuals in PO are sorted based on Pareto-dominance sorting scheme proposed in [38] (Line 2 in Alg. 3). Then, the ranking is updated according to Alg. 4 where the ψ-quasi equal feature subsets will be pulled to the same front (Line 11 in Alg. 4). For example, in Fig. 1, feature subset S_2 from the second front will be dragged into the first front which has the same front value as S_1. Likewise, feature subsets S_4 and S_5 or feature subsets S_6 and S_7 will also sit at the same front, respectively. Noted that there may be a large number of ψ-quasi equal feature subsets in one group $Qe(i)$, e.g., S_1-S_7 in Fig. 3. If the algorithm does not limit the number of ψ-quasi equal feature subsets and keep all of them during evolution, that will reduce the diversity of solutions in the objective space and may cause an algorithm to convergence prematurely. For example, suppose that five solutions need to be chosen from Fig. 3 to form the population to the next generation, the algorithm may select solution S_1-S_5. However, the other solutions with lower classification error rate, e.g., S_8-S_{10}, are lost. To alleviate the above impact, Eq. (5) is used to limit the number (termed ln) of ψ-quasi equal feature subsets in each group.

$$ln = |PO| // \text{unique\_number}(f_2(\mathcal{W}_1)) \tag{5}$$

where $\text{unique\_number}(f_2(\mathcal{W}_1))$ means the number of unique size of the solutions in the first front. $|PO|$ means the number of solutions in PO, and $//$ means floor division.

In Fig. 3, $\text{unique\_number}(f_2(\mathcal{W}_1)) = 4$ and $|PO| = 10$, since solutions S_1-S_{10} in the relaxed first front have four unique FR values. Using Eq. (5), we can get $ln = 10//4 = 2$.

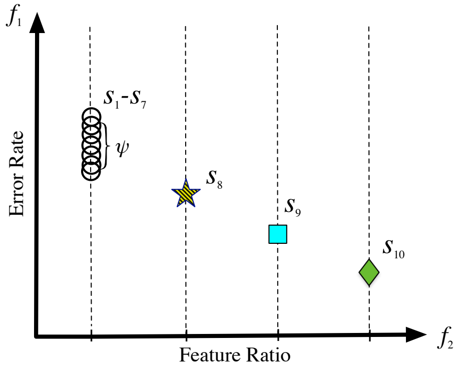

Fig. 3. An example of the distribution of the objective values of 10 solutions. Based on the ***Definition 1***, S_1 to S_7 are ψ-quasi equal feature subsets.

Therefore, NMDE will pick two solutions from S_1-S_7 with S_8-S_{10} form a new population for the next generation.

After getting the updated fronts ($\mathcal{W}_1$-$\mathcal{W}_k$) during evolution from Alg. 4, the crowding distance of individuals in $\mathcal{W}_m$ (the m-th front), $m \in [1, k]$, will be calculated. The stimulus condition is the total number of solutions in the first $m-1$ fronts is smaller than the population size and this number in the first $m+1$ fronts is larger than the population size. Next, the solutions in $\mathcal{W}_m$ will be ordered in a descending order (Line 4 in Alg. 3) based on their crowding distances. Meanwhile, the individuals in the first $m-1$ fronts are put in a temporary set $\mathcal{T}$. Details of calculating the crowding distance of an individual are shown as follows.

An aggregated function to calculate the crowding estimation (i.e., C_i) of individual $\vec{x}_i$, which considers the crowding estimation both in the objective space (i.e., C_i^o) and in the solution space (i.e., C_i^s), is shown Eq. (6).

$$C_i = \begin{cases} \max\{C_i^o, C_i^s\} \text{ if } C_i^o > C_{\text{avg}}^o \textit{ or } C_i^s > C_{\text{avg}}^s \\ \min\{C_i^o, C_i^s\} \qquad\qquad\qquad\qquad \text{otherwise} \end{cases} \tag{6}$$

where c_{avg}^o and c_{avg}^s represent the average crowding distances in the objective space and the solution space of the solutions in $\mathcal{W}_m$, respectively. Note that C_i^o and C_i^s in Eq. (6) are normalized. The details of computing the two main components, C_i^o and C_i^s, are given as follows.

1) The determination of C_i^o: The crowding distance of individual $\vec{x}_i$ in the objective space is termed as C_i^o. Since traditional crowding distance [38], independently calculated in each Pareto rank, may not truly reflect the crowding degree in the objective space [26], the proposed strategy considers the already selected individuals in earlier fronts when calculating C_i^o. Furthermore, the ψ-quasi equal feature subsets are set to have the same crowding distance in the objective space.

$$C_i^o = \begin{cases} 1 & \text{if } \vec{x}_i \textit{ is a boundary point} \\ \sum_{j=1}^{obj} \frac{f_j(\vec{x}_{i+1}) - f_j(\vec{x}_{i-1})}{obj*(f_j^{\max} - f_j^{\min})} & \text{otherwise} \end{cases} \tag{7}$$

where $\vec{x}_{i-1}$ and $\vec{x}_{i+1}$ are the left neighbor and the right neighbor of $\vec{x}_i$ in $\mathcal{T}$, respectively; $f_j(\vec{x}_{i-1})$ and $f_j(\vec{x}_{i+1})$ are the j-th objective value of the individual $\vec{x}_{i-1}$ and $\vec{x}_{i+1}$, respectively. Meanwhile, $f_j^{\max}$ and $f_j^{\min}$ are the upper boundary and the lower boundary of the j-th objective value of all the solutions in $\mathcal{T}$, respectively. In this work, $obj = 2$.

Algorithm 5: NMDE

```
Input: $NP$, $Max\_FEs$, and $\psi$
Output: $P$
begin
    Randomly initialize population with size $NP$, and set current
      fitness evaluations $fe = 0$;
    Calculate the objective values of the generated individuals,
      $fe = fe + NP$,
    while $fe < Max\_FEs$ do
        for $i = 1, \ldots, NP$ do
            Find the niche of the current individual $\vec{x}_i$,
            Generate a new individual via Eq. (2) and crossover,
        end
        Calculate the objective values of the new individuals,
        Perform subset repairing scheme, and update $fe$,
        Perform environmental selection, and get $P$,
    end
    Output the solutions in the first front and the $\psi$-quasi equal
      feature subsets from $P$.
end
```

Eq. (7) shows that the crowding distance of the solutions whose objective values are the lowest or the largest in $\mathcal{T}$ is set to 1 rather than ∞. This boundary-point metric assignment is followed by the study in [16], which is to ensure that C_i^o and C_i^s can be compared or combined. For instance, in Fig. 1, when calculating the crowding distance of S_7 in the objective space (termed C_7^o), its left neighbor and the right neighbor will be determined from $\mathcal{T}$: $\{S_1, S_2, S_4, S_5, S_8, S_9, S_{10}\}$. After getting C_7^o by Eq. (7), S_6 will have the same crowding distance as S_7 in the objective space, i.e., $C_6^o = C_7^o$.

2) The determination of C_i^s: The second component is the crowding distance of individual $\vec{x}_i$ in the solution space, named C_i^s. C_i^s is the average of the Hamming distances (d) between individual $\vec{x}_i$ to the solution in its niche ($\mathcal{N}_i$), which is shown in Eq. (8).

$$C_i^s = \frac{1}{|\mathcal{N}_i|} \sum_{j=1}^{|\mathcal{N}_i|} d_{ij} \tag{8}$$

where j is the order number of the individuals in the niche of $\vec{x}_i$. The number $j = 1$ for the nearest neighbor of $\vec{x}_i$, and $j = |\mathcal{N}_i|$ for the farthest neighbor.

Eq. (8) shows a measure of sparseness at a point, which is often used in novelty search. If the average distance to a given point's neighbors is large, then it is in a sparse area, i.e., this solution is novel. Otherwise, it is in a dense region. If C_i^s is larger, the individual should have a higher chance to enter into the next generation.

G. Overall Algorithm

This section gives the complete NMDE algorithm in Alg. 5. The initialization and the crossover operators are inherited from the traditional DE. The mutation, the subset repairing, and the environmental selection operators have been introduced as above. The maximal number of fitness evaluations is termed Max_FEs. When Max_FEs is reached, NMDE stops and returns the solutions in the first front and the ψ-quasi equal feature subsets as the result.

TABLE I: The information of datasets

Number	Dataset	# Features	# Classes	# Instances
1	Zoo	16	7	101
2	SPECT	22	2	267
3	WBCD	30	2	569
4	Ionosphere	34	2	351
5	Sonar	60	2	208
6	Movement	90	15	360
7	Hillvally	100	2	606
8	Musk1	166	2	476
9	Multiple (pix)	240	10	2,000
10	Madelon	500	2	4,400
11	CNAE	856	9	1,080
12	AD	1,558	2	3,279
13	SRBCT	2,308	4	83
14	Leukemia1	5,327	3	72
15	DLBCL	7,050	2	77
16	Leukemia2	11,225	3	72
17	11Tumor	12,533	11	174
18	Lung Cancer	12,600	5	203
19	14Tumor	15,009	26	308

IV. Experiment Design

A. Benchmark Techniques

Six EMO methods are chosen for comparisons with NMDE, i.e., Omni-optimizer [55], DN-NSGAII [56], MO_Ring_PSO_SCD [16], MMODE_ICD [26], SparseEA [49], and GF-NSGAII [46]. Among them, Omni-optimizer, DN-NSGAII, MO_Ring_PSO_SCD, and MMODE_ICD are four multimodal multi-objective optimization algorithms, which are used for finding multiple optimal solutions of a multimodal problem. SparseEA is a sparse-based multi-objective feature selection method, and GF-NSGAII is a grid-dominance based multi-objective feature selection method. All the compared algorithms are implemented using codes from the corresponding authors.

B. Datasets and Parameter Settings

There are nineteen datasets used in the experiments, and the detailed information of those datasets can be seen in Table I. Those datasets are from different fields, e.g., physic/chemistry (Sonar), health (WBCD), and hand written recognition (Multiple). The datasets 13-19 are from the biomedical domain, which include thousands of features but tens of instances/samples (SRBCT and DLBCL) and up to $10,000$ features and a few hundreds of instances/samples (Lung Cancer and 11Tumor). Those 7 gene expression datasets pose a huge challenge to classification and feature learning algorithms, due to the high dimensionality but a small sample size.

On each dataset, each algorithm runs 30 times independently on each dataset, i.e., the random seed is set the same for all the algorithms but changes by runs. The specific parameters of the compared algorithms are from the corresponding literatures [55], [56], [16], [26], [49], [46]. The population size is set to the number of features in a dataset but restricted to 300 in order to balance the diversity and efficiency. The maximal number (Max_FEs) of objective function evaluations is set to 100 times of NP for all the datasets. The objective function evaluations are consumed to calculate the classification errors rates of feature subsets.

In terms of classification, each dataset is randomly separated into a training set and a test set with the proportions of nearly 70% and 30%, respectively. During the training process, KNN with 5-fold cross-validation on the training set is utilized to calculate the classification error rate to avoid the feature selection bias, and the K values in KNN is set to five so as to balance the accuracy and efficiency.

For NMDE, F in Eq. (2) is set to 0.5, and the crossover rate (CR) is set to 0.5 [26]. For the parameter ψ in NMDE, ψ at $e-02$, $e-04$, $e-10$, and 0 are adopted in the experiments, respectively. One note is that this paper mainly discuss the experimental results at $\psi = e-04$. The reasons are as follows. In Table I, the largest number of instances is $4,400$ on the Madelon dataset, and the number of training instances on the Madelon dataset is $4,400 \times 0.7 = 3,080$. Obviously, $e-04$ is smaller than $1/3,080 \approx 3.2 * e-04$, thus ψ at $e-04$ does not relax the accuracy so much. The threshold θ is set to 0.6, as suggested in [3], [46].

C. Performance Indicators

For the analysis of the experimental results, six performance indicators are employed to compare different feature selection methods, i.e., HV, IGD, the number of selected features ($Size$), the classification accuracy, the number of feature subsets with the highest training accuracy (Num), and squared cosine redundancy rate (RED). Among them, HV and IGD are two popular comprehensive indicators to measure the convergence and the diversity of the feature subsets obtained from a multi-objective method. Another indicator, RED, is widely used to measure the redundancy rate of a feature subset (e.g., S) [57]–[59], which is defined as:

$$RED(S) = \frac{1}{d * (d-1)} \sum_{F_i, F_j \in S, i \neq j} \cos^2(F_i, F_j) \tag{9}$$

where F_i and F_j mean the i-th feature and the j-th feature in the d selected features in S, respectively. The measure assesses the average similarity among all the selected feature pairs and produces values between 0 and 1. A large value indicates high redundancy in S, thus a lower RED value means a better feature subset with less redundancy.

V. Results

Tables II-III show the average HV and IGD results of the seven algorithms on the 19 datasets. The reference point of HV is set to $(1, 1)$. For calculating the IGD, unlike other multi-objective and/or multimodal benchmark problems, e.g., MMF1 to MMF8 [16], feature selection is a discrete problem and the *true* PF is unknown. Therefore, all the output feature subsets from the seven algorithms are collected and merged into one set. Then, all feature subsets in the set are ranked by the non-dominated sorting. Next, only the first PF is kept as the reference point of IGD. The higher the HV values or the smaller the IGD values, the better the feature subsets. In Tables II-III, the highest HV and the lowest IGD values obtained on each dataset are in bold. The signs '$\uparrow$', '$\downarrow$', and 'o' indicate that the corresponding benchmark algorithm is significantly better than, worse than or has no

TABLE II: The average HV results on the test sets (the larger the better).

Dataset	Omni-optimizer	DN-NSGAII	MO_Ring_PSO_SCD	MMODE_ICD	SparseEA	GF-NSGAII	NMDE
Zoo	7.877e-01 ±4.540e-02 ↓	7.734e-01 ±5.100e-02 ↓	**8.476e-01** ±1.130e-02 ↑	8.350e-01 ±1.180e-02 o	8.370e-01 ±8.400e-03 o	8.275e-01 ±1.260e-02 ↓	8.396e-01 ±7.600e-03
SPECT	6.946e-01 ±4.730e-02 ↓	6.835e-01 ±3.560e-02 ↓	7.593e-01 ±1.320e-02 o	7.650e-01 ±0.000e+00 o	**7.668e-01** ±4.700e-03 o	7.666e-01 ±3.700e-03 o	7.665e-01 ±1.200e-02
WBCD	8.030e-01 ±3.620e-02 ↓	8.131e-01 ±3.690e-02 ↓	9.134e-01 ±8.900e-03 o	9.071e-01 ±1.120e-02 ↓	9.099e-01 ±1.130e-02 ↓	9.117e-01 ±4.200e-03 ↓	**9.166e-01** ±5.000e-03
Ionosphere	7.461e-01 ±5.000e-02 ↓	7.325e-01 ±4.360e-02 ↓	**8.933e-01** ±1.910e-02 o	8.880e-01 ±2.700e-02 o	8.934e-01 ±3.160e-02 o	8.857e-01 ±1.800e-02 o	8.854e-01 ±2.030e-02
Sonar	7.127e-01 ±3.190e-02 ↓	7.013e-01 ±3.220e-02 ↓	8.091e-01 ±3.010e-02 ↓	8.468e-01 ±2.950e-02 o	**8.473e-01** ±3.050e-02 o	8.241e-01 ±3.710e-02 ↓	8.450e-01 ±3.030e-02
Movement	6.442e-01 ±1.840e-02 ↓	6.552e-01 ±2.020e-02 ↓	7.109e-01 ±1.300e-02 ↓	7.536e-01 ±2.600e-02 ↓	7.758e-01 ±1.660e-02 ↓	7.438e-01 ±2.170e-02 ↓	**7.856e-01** ±1.600e-02
Hillvally	4.755e-01 ±1.530e-02 ↓	4.714e-01 ±1.660e-02 ↓	5.262e-01 ±1.190e-02 ↓	5.815e-01 ±1.430e-02 ↓	5.928e-01 ±7.500e-03 ↓	5.932e-01 ±1.050e-02 ↓	**6.035e-01** ±8.700e-03
Musk1	7.686e-01 ±1.990e-02 ↓	7.597e-01 ±1.630e-02 ↓	8.132e-01 ±1.020e-02 ↓	9.105e-01 ±1.930e-02 ↓	9.709e-01 ±3.100e-03 ↓	9.579e-01 ±7.000e-03 ↓	**9.766e-01** ±4.200e-03
Multiple	7.526e-01 ±1.440e-02 ↓	7.451e-01 ±1.040e-02 ↓	7.686e-01 ±9.500e-03 ↓	8.299e-01 ±1.530e-02 ↓	9.514e-01 ±2.000e-03 o	8.925e-01 ±1.450e-02 ↓	**9.517e-01** ±3.300e-03
Madelon	5.892e-01 ±1.470e-02 ↓	5.886e-01 ±1.320e-02 ↓	6.157e-01 ±6.400e-03 ↓	7.925e-01 ±2.420e-02 ↓	**8.997e-01** ±5.900e-03 o	8.896e-01 ±4.000e-03 ↓	8.985e-01 ±6.500e-03
CNAE	5.931e-01 ±1.590e-02 ↓	5.905e-01 ±1.490e-02 ↓	6.167e-01 ±9.200e-03 ↓	7.070e-01 ±1.380e-02 ↓	**8.656e-01** ±8.700e-03 ↑	8.290e-01 ±1.010e-02 ↓	8.564e-01 ±1.360e-02
AD	6.807e-01 ±9.800e-03 ↓	6.799e-01 ±5.900e-03 ↓	6.897e-01 ±3.400e-03 ↓	7.476e-01 ±1.250e-02 ↓	**9.795e-01** ±2.300e-03 o	8.988e-01 ±1.240e-02 ↓	9.786e-01 ±1.500e-03
SRBCT	6.623e-01 ±1.910e-02 ↓	6.629e-01 ±1.950e-02 ↓	6.922e-01 ±1.050e-02 ↓	7.356e-01 ±1.670e-02 ↓	9.180e-01 ±3.680e-02 ↓	8.942e-01 ±1.260e-02 ↓	**9.833e-01** ±2.600e-02
Leukemia1	5.640e-01 ±3.160e-02 ↓	5.465e-01 ±3.590e-02 ↓	5.949e-01 ±2.610e-02 ↓	6.195e-01 ±3.370e-02 ↓	8.786e-01 ±6.120e-02 o	6.998e-01 ±5.920e-02 ↓	**8.965e-01** ±5.910e-02
DLBCL	6.552e-01 ±1.660e-02 ↓	6.579e-01 ±1.260e-02 ↓	6.830e-01 ±7.500e-03 ↓	7.070e-01 ±9.000e-03 ↓	**9.618e-01** ±4.580e-02 o	7.896e-01 ±1.540e-02 ↓	9.405e-01 ±4.220e-02
Leukemia2	5.970e-01 ±1.620e-02 ↓	5.991e-01 ±1.250e-02 ↓	6.234e-01 ±1.240e-02 ↓	6.369e-01 ±1.340e-02 ↓	9.297e-01 ±3.870e-02 o	7.197e-01 ±1.940e-02 ↓	**9.403e-01** ±3.430e-02
11Tumor	5.265e-01 ±1.960e-02 ↓	5.291e-01 ±1.840e-02 ↓	5.586e-01 ±1.380e-02 ↓	5.688e-01 ±1.580e-02 ↓	7.761e-01 ±3.880e-02 ↓	6.357e-01 ±2.450e-02 ↓	**7.989e-01** ±4.130e-02
Lung Cancer	6.234e-01 ±9.000e-03 ↓	6.221e-01 ±8.300e-03 ↓	6.415e-01 ±8.600e-03 ↓	6.660e-01 ±9.900e-03 ↓	9.033e-01 ±3.090e-02 ↓	7.201e-01 ±1.130e-02 ↓	**9.269e-01** ±2.210e-02
14Tumor	3.474e-01 ±1.170e-02 ↓	3.511e-01 ±1.010e-02 ↓	3.635e-01 ±1.200e-02 ↓	3.757e-01 ±9.200e-03 ↓	4.987e-01 ±1.450e-02 ↓	3.992e-01 ±1.150e-02 ↓	**5.323e-01** ±2.510e-02
Rank	6.29 (0/0/19)	6.63 (0/0/19)	4.58 (1/3/15)	3.68 (0/4/15)	2.25 (1/10/8)	3.10 (0/2/17)	1.40

significant difference from NMDE, respectively. The Wilcoxon significance test with a level of 0.05 is used. Moreover, the Freidman test is employed to give the relative performance ranking among the seven algorithms.

A. Overall Analysis

The results clearly show that NMDE achieves the best among the seven algorithms in terms of the test HV and IGD results.

As shown in Table II among the 114 comparisons with the other six algorithms, there are only two losses of NMDE on HV. One significant loss of NMDE happens against MO_Ring_PSO_SCD on the Zoo test set, and another loss happens against SparseEA on the CNAE test set, but NMDE still takes the second-best on the Zoo and CNAE datasets among the seven algorithms. The IGD performance shown in Table III reveals a similar pattern to HV. There are six losses out of the 114 comparisons of NMDE on IGD. Specifically, NMDE loses to SparseEA on three datasets, i.e., the SPECT, Madelon, and CNAE datasets. In addition, both MO_Ring_PSO_SCD and MMODE_ICD achieve better IGD performance than NMDE only on the SPECT dataset. The final loss of NMDE happens in the comparisons with MO_Ring_PSO_SCD on the Ionosphere dataset.

For an intuitive analysis, the distributions of the first non-dominated fronts from the median run obtained by each algorithm are shown in Fig. 4 where the first and the second rows show the results on the training set and test set, respectively. Noted that all the feature subsets in the second row come from the first row. In each sub-figure, the two numbers inside the brackets indicate the number of the original features and the training or the test error rate of using all features in a dataset. Four datasets, i.e., the Sonar, SRBCT, Leukemia1, and 11Tumor datasets, are chosen since their PF distributions from the seven algorithms are easier to be distinguished visually.

Fig. 4 shows that the median PFs evolved by NMDE are usually more diverse than the ones by the compared six algorithms both on the training and test sets. In addition, it is obvious although some feature subsets obtained from the seven algorithms have the same training and/or testing classification accuracy, the feature subsets from NMDE includes fewer features. For example, on the SRBCT dataset, although different feature subsets from the seven algorithms can achieve both the highest training and testing classification accuracy (i.e., nearly 100%), the feature subsets from NMDE only contains less than 4% of the original features, but this number for Omni-optimizer, DN-NSGAII, MO_Ring_PSO_SCD, and MMODE_ICD is over 30%. For GF-NSGAII, this number is

TABLE III: The average IGD results on the test sets (the smaller the better).

Dataset	Omni-optimizer	DN-NSGAII	MO_Ring_PSO_SCD	MMODE_ICD	SparseEA	GF-NSGAII	NMDE
Zoo	1.069e-01 ±2.830e-02 ↓	1.143e-01 ±3.000e-02 ↓	**5.130e-02** ±1.140e-02 o	6.140e-02 ±1.180e-02 ↓	6.400e-02 ±9.100e-03 ↓	6.860e-02 ±8.800e-03 ↓	5.400e-02 ±9.600e-03
SPECT	6.850e-02 ±3.070e-02 ↓	7.700e-02 ±3.410e-02 ↓	3.060e-02 ±1.210e-02 ↑	**2.540e-02** ±7.400e-03 ↑	2.660e-02 ±1.110e-02 ↑	4.420e-02 ±1.570e-02 o	4.010e-02 ±1.070e-02
WBCD	1.093e-01 ±3.220e-02 ↓	9.740e-02 ±3.540e-02 ↓	2.180e-02 ±6.800e-03 ↓	2.290e-02 ±6.900e-03 ↓	1.980e-02 ±7.000e-03 ↓	1.960e-02 ±4.000e-03 ↓	**1.550e-02** ±5.500e-03
Ionosphere	1.293e-01 ±3.780e-02 ↓	1.442e-01 ±3.650e-02 ↓	3.300e-02 ±1.190e-02 ↑	3.840e-02 ±1.630e-02 o	**3.050e-02** ±2.080e-02 o	3.870e-02 ±9.800e-03 o	3.950e-02 ±1.100e-02
Sonar	1.608e-01 ±2.730e-02 ↓	1.679e-01 ±2.390e-02 ↓	8.880e-02 ±1.440e-02 ↓	6.990e-02 ±1.780e-02 o	6.990e-02 ±1.820e-02 o	7.610e-02 ±2.460e-02 o	**6.770e-02** ±1.980e-02
Movement	2.273e-01 ±1.320e-02 ↓	2.173e-01 ±2.000e-02 ↓	1.616e-01 ±1.390e-02 ↓	1.048e-01 ±4.100e-02 ↓	5.320e-02 ±8.900e-03 ↓	7.190e-02 ±1.890e-02 ↓	**4.660e-02** ±9.500e-03
Hillvally	1.672e-01 ±2.290e-02 ↓	1.684e-01 ±2.130e-02 ↓	9.780e-02 ±1.200e-02 ↓	3.830e-02 ±1.540e-02 ↓	3.050e-02 ±5.000e-03 ↓	3.400e-02 ±4.600e-03 ↓	**2.760e-02** ±4.600e-03
Musk1	1.735e-01 ±2.170e-02 ↓	1.817e-01 ±1.700e-02 ↓	1.314e-01 ±9.700e-03 ↓	4.770e-02 ±1.320e-02 ↓	1.640e-02 ±3.600e-03 ↓	3.360e-02 ±5.200e-03 ↓	**1.300e-02** ±3.200e-03
Multiple	2.342e-01 ±1.360e-02 ↓	2.419e-01 ±1.050e-02 ↓	2.192e-01 ±9.500e-03 ↓	1.605e-01 ±1.380e-02 ↓	**1.780e-02** ±2.800e-03 o	9.820e-02 ±1.220e-02 ↓	1.820e-02 ±3.000e-03
Madelon	3.083e-01 ±1.210e-02 ↓	3.114e-01 ±1.270e-02 ↓	2.916e-01 ±5.800e-03 ↓	1.564e-01 ±2.070e-02 ↓	**1.190e-02** ±3.900e-03 ↑	1.270e-01 ±5.600e-03 ↓	4.370e-02 ±3.080e-02
CNAE	3.043e-01 ±1.080e-02 ↓	3.066e-01 ±7.500e-03 ↓	2.998e-01 ±7.500e-03 ↓	2.462e-01 ±1.250e-02 ↓	**2.590e-02** ±5.600e-03 ↑	1.717e-01 ±6.400e-03 ↓	5.670e-02 ±1.510e-02
AD	3.004e-01 ±1.009e-02 ↓	3.021e-01 ±6.100e-03 ↓	2.924e-01 ±3.530e-03 ↓	2.306e-01 ±1.319e-02 ↓	2.587e-02 ±4.760e-03 ↓	8.876e-02 ±9.780e-02 ↓	**2.088e-02** ±5.600e-03
SRBCT	3.819e-01 ±1.080e-02 ↓	3.837e-01 ±1.050e-02 ↓	3.684e-01 ±6.500e-03 ↓	3.401e-01 ±1.450e-02 ↓	**4.470e-02** ±1.610e-02 o	2.232e-01 ±1.020e-02 ↓	4.680e-02 ±2.370e-02
Leukemia1	3.952e-01 ±1.470e-02 ↓	4.028e-01 ±2.220e-02 ↓	3.694e-01 ±1.070e-02 ↓	3.412e-01 ±1.560e-02 ↓	8.590e-02 ±4.530e-02 o	2.344e-01 ±3.180e-02 ↓	**6.620e-02** ±4.570e-02
DLBCL	3.705e-01 ±4.200e-03 ↓	3.706e-01 ±4.400e-03 ↓	3.493e-01 ±2.400e-03 ↓	3.227e-01 ±7.400e-03 ↓	**3.130e-02** ±3.750e-02 o	2.266e-01 ±4.900e-03 ↓	4.040e-02 ±3.260e-02
Leukemia2	3.819e-01 ±6.800e-03 ↓	3.831e-01 ±4.900e-03 ↓	3.743e-01 ±5.130e-03 ↓	3.442e-01 ±7.300e-03 ↓	4.190e-02 ±3.100e-02 o	2.552e-01 ±7.500e-03 ↓	**3.620e-02** ±2.310e-02
11Tumor	3.884e-01 ±3.500e-03 ↓	3.887e-01 ±3.300e-03 ↓	3.591e-01 ±3.500e-03 ↓	3.556e-01 ±6.100e-03 ↓	7.170e-02 ±2.460e-02 o	3.041e-01 ±7.300e-03 ↓	**6.320e-02** ±1.790e-02
Lung Cancer	3.157e-01 ±3.300e-03 ↓	3.160e-01 ±3.100e-03 ↓	3.132e-01 ±3.200e-03 ↓	2.792e-01 ±5.600e-03 ↓	9.760e-02 ±1.560e-02 ↓	2.328e-01 ±2.300e-03 ↓	**8.280e-02** ±1.000e-02
14Tumor	3.647e-01 ±2.900e-03 ↓	3.643e-01 ±2.100e-03 ↓	3.524e-01 ±2.620e-03 ↓	3.332e-01 ±6.200e-03 ↓	5.020e-02 ±1.190e-02 ↓	2.859e-01 ±3.600e-03 ↓	**3.990e-02** ±7.700e-03
Rank	6.21 (0/0/19)	6.79 (0/0/19)	4.50 (2/1/16)	3.89 (1/2/16)	2.14 (3/8/8)	3.07 (0/3/16)	1.56

over 10%. Although the number of selected features between SparseEA and NMDE is close, the feature subsets from NMDE have better distribution than that from SparseEA in the objective space.

In summary, NMDE wins 187, draws 33 and losses 8 out of the 228 comparisons in terms of the test HV and IGD results. The results suggest that the non-dominated feature subsets of NMDE are generally more diverse (i.e., having more non-dominated solutions) and converge better (i.e., with fewer selected features and lower classification error rates) than those of the compared six multi-objective feature selection algorithms, i.e., Omni-optimizer, DN-NSGAII, MO_Ring_PSO_SCD, MMODE_ICD, GF-NSGAII, and SparseEA.

B. Test Performance Analysis

Niching-based EMO methods can get a set of feature subsets as shown in Fig. 6. The solutions which are ψ-quasi equal feature subset to the feature subset with the lowest training error rate from one method are probably the most valuable solutions for a decision-maker. Therefore, these feature subsets, e.g., the last three feature subsets in Fig. 6, will be kept in a set $same\_lowest\_error$. The test performance of the feature subsets in $same\_lowest\_error$ will be reported. Specifically, the size of $same\_lowest\_error$, i.e., the number of feature subsets in $same\_lowest\_error$, is termed Num. The average number of selected features of the feature subsets in $same\_lowest\_error$ is termed $Size$. In addition, the average test accuracy of the solutions in $same\_lowest\_error$ is also reported in Fig. 5. Eight datasets including the WBCD, Sonar, Musk1, Madelon, SRBCT, Leukemia1, 11Tumor, and 14Tumor datasets are chosen as representative to show the performance in Fig. 5. The results on the remaining 11 datasets reveal similar patterns.

As shown in Fig. 5, NMDE has the largest Num values than the six algorithms on almost all the 19 datasets. Only on the 11Tumor dataset, NMDE does not get the largest Num values. The highest improvement can be seen on the Leukemia1 dataset, NMDE can find over 35 different feature subsets with the same training accuracy, while this number on the other methods is less than 3. More importantly, those different feature subsets from NMDE on the Leukemia1 dataset include less than 100 features but with the highest test accuracy. Furthermore, NMDE also achieves the highest accuracy on the WBCD and Sonar datasets with fewer features included than that from Omni-optimizer, DN-NSGAII, and MO_Ring_PSO_SCD. On the Musk1, Madelon, SRBCT, and 14Tumor datasets, NMDE can achieve the second-best of

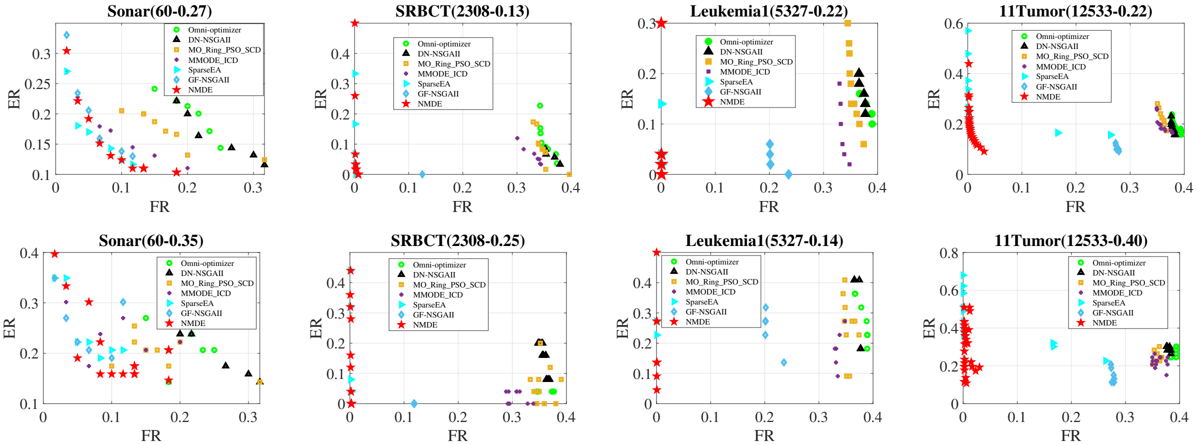

Fig. 4. The obtained PFs of seven algorithms on the training and the test sets (upper row for training results and bottom row for test results).

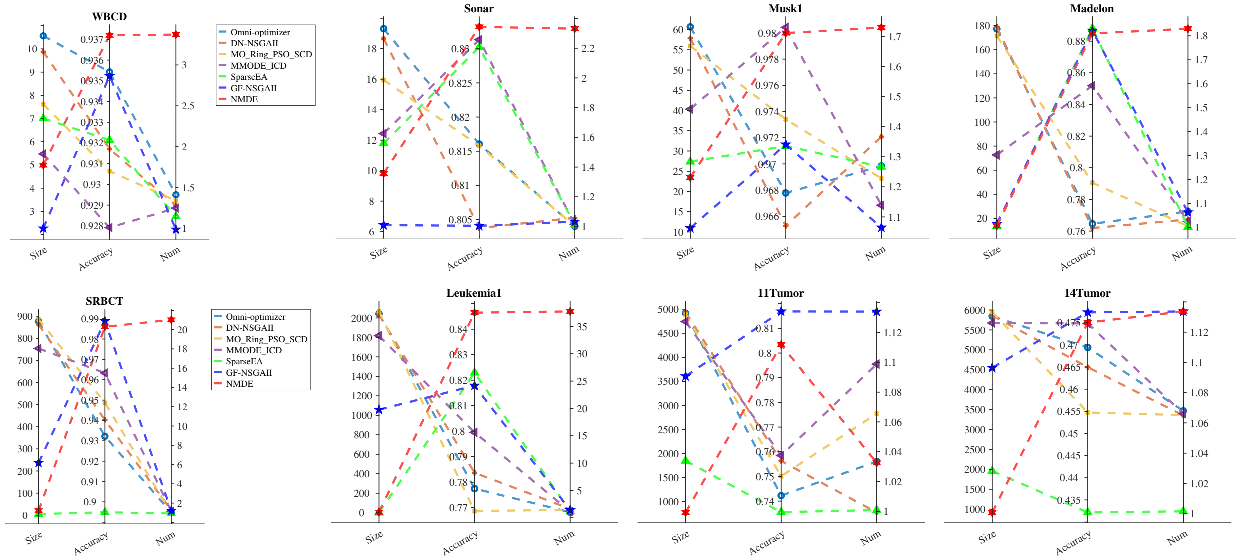

Fig. 5. The average test results of the feature subsets with the lowest training error rate from the seven algorithms on the 8 datasets. The average number of features is termed $Size$, the average test accuracy is termed $Accuracy$, and the average number of feature subsets with the lowest training error rate is termed Num.

accuracy among the seven methods. On the 11Tumor dataset, the superiority of NMDE on the accuracy is slightly decreases but still have a significant advantage on $Size$ results.

The results show that NMDE can generally achieve promising and excellent classification performance for feature selection in classification. More importantly, NMDE can find more different feature subsets with the same training performance and promising test performance.

C. Major Component Contribution Analysis

NMDE has three major components: the developed mutation operator, the subset repairing mechanism, and the environmental selection strategy. Eight algorithms are made to test the performance of the three components, and their convergence performance, i.e., HV and IGD results, are shown in Tables S.I and S.II in the Online Supplementary Materials.

1) Mutation Operator: To test the performance of the proposed mutation operator, five baseline algorithms using five commonly used mutation operators are made. Then, the five methods are compared with a new DE-based EMO method (named MODE-NM) with the proposed mutation operator. The only difference among the six methods is they employ different mutation operators. The HV and IGD results of the six methods are included in Tables S.I and S.II. The average training HV plots with generations of the six methods are shown in Fig. S.1 in the Online Supplementary Materials.

TABLE IV: The avarage RED results between NMDE-N (NMDE without the proposed subset repairing mechanism) and NMDE.

	NMDE-N	NMDE		NMDE-N	NMDE
Zoo	**1.722e-01** ±2.085e-02 o	1.728e-01 ±2.145e-02	CNAE	4.922e-01 ±9.662e-04 ↓	**4.753e-01** ±5.289e-04
SPECT	5.962e-02 ±4.082e-02 o	**5.931e-02** ±3.308e-02	AD	4.924e-01 ±7.934e-04 ↓	**4.235e-01** ±1.506e-04
WBCD	2.138e-02 ±1.553e-03 ↓	**3.904e-03** ±1.715e-04	SRBCT	3.675e-01 ±1.106e-02 ↓	**2.701e-01** ±2.351e-02
Ionosphere	2.173e-01 ±4.784e-02 o	**2.091e-01** ±4.455e-02	Leukemia1	2.770e-01 ±3.255e-02 ↓	**1.855e-01** ±2.697e-02
Sonar	2.684e-02 ±3.370e-03 o	**2.616e-02** ±2.250e-03	DLBCL	4.705e-01 ±1.650e-02 ↓	**2.002e-01** ±1.679e-02
Movement	4.507e-03 ±1.020e-03 ↓	**3.453e-03** ±4.641e-04	Leukemia2	3.387e-01 ±1.752e-02 ↓	**1.465e-01** ±1.071e-02
Hillvally	1.195e-05 ±2.911e-06 o	**1.129e-05** ±2.288e-06	11Tumor	4.169e-01 ±1.093e-02 ↓	**3.896e-01** ±1.754e-02
Musk1	**5.505e-01** ±3.928e-02 o	5.515e-01 ±2.352e-02	Lung Cancer	3.009e-01 ±1.043e-02 ↓	**2.728e-01** ±1.252e-02
Multiple	1.050e-01 ±6.506e-03 ↓	**9.740e-02** ±6.157e-03	14Tumor	3.637e-01 ±7.526e-03 ↓	**3.495e-01** ±4.154e-03
Madelon	2.019e-04 ±5.223e-05 ↓	**4.530e-05** ±1.612e-05			

As shown in Fig. S.I, MODE-NM achieves the fast conver-

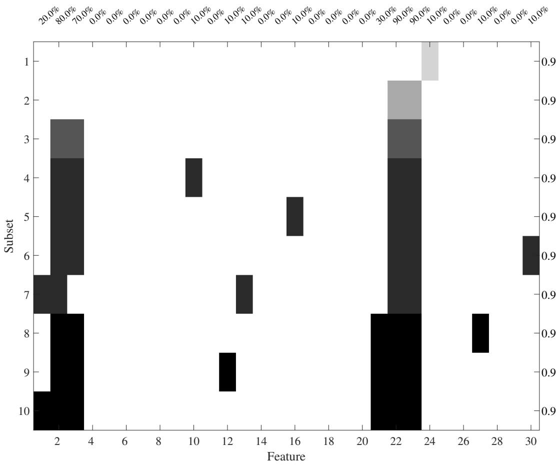

Fig. 6. Frequency matrix from NMDE on the WBCD dataset. 'Feature' means one of the 30 original features in the WBCD dataset, and 'Subset' means one of the obtained feature subsets from NMDE. The training accuracy using each feature subset is shown on the right side, and the selected frequency of each feature is shown on the upper side. The square (i, j) will be highlighted if the j-th feature is selected in the i-th feature subset. The color becomes darker as the subset size increases, i.e., the feature subsets are ranked based on their sizes.

gence and the largest HV values than MODE-rand1, MODE-best1, MODE-current-rand1, MODE-rand2, and MODE-best2 on almost all the training sets. This trend is particularly obvious as the number of features increases in a dataset. More importantly, MODE-NM achieves significantly better rankings than MODE-rand1, MODE-best1, MODE-current-rand1, MODE-rand2, and MODE-best2 on both HV and IGD results in Tables S.I and S.II. The superiority of MODE-NM in IGD is more obvious than that in HV.

The results show that the proposed mutation operator accelerates the convergence of the algorithm and obtains feature subsets with better HV and IGD performance.

2) Environmental Selection Strategy: To explore the effect of the proposed environmental selection strategy, another variant algorithm of NMDE, i.e., NMDE-N, is formed. NMDE-N is based on MODE-NM and incorporates the proposed environmental selection strategy. The results between MODE-NM and NMDE-N in Tables S.I and S.II show that the proposed environmental selection strategy can help the proposed NMDE algorithm achieve better HV and IGD results. More importantly, in Section V.B, the results show that the proposed environmental selection strategy can help NMDE find more ψ-quasi equal feature subsets.

3) Subset Repairing Strategy: NMDE is formed using all the three major components. The performance of the proposed subset repairing mechanism can be seen from the comparison between NMDE-N and NMDE in Tables S.I and S.II in the Online Supplementary Materials. The results show that the proposed subset repairing mechanism by modifying the ψ-quasi equal feature subsets during evolution achieve higher HV and/or lower IGD values in most of the 19 datasets. Table IV gives the RED results between NMDE and NMDE-N. As shown in Table IV, the feature subsets obtained from NMDE show a significantly lower redundancy rate than that from NMDE-N on 13 out of the 19 datasets. The highest improvement is seen on the DLBCL dataset with 0.27 decrease on average. On the remaining 6 datasets including the Zoo, SPECT, Ionosphere, Sonar, Hillvally, and Musk1 datasets, the RED values between NMDE and NMDE-N have no significant difference. The results suggest that the proposed subset repairing mechanism can help NMDE obtain feature subsets with lower redundant rate.

In summary, all of the three major components of NMDE can improve the performance of an algorithm and will generally make greater contributions if combined.

D. Further Analysis on Selected Features

This sub-section analyzes the different trade-offs between the number of selected features and the classification accuracy. This helps with the analysis and understanding of the relevant features and the relative importance of each selected feature. The results of NMDE on the WBCD dataset are shown in Fig. 6 since the WBCD dataset only has 30 features that are easy to display. In Fig. 6, each column (Feature) represents one of the 30 original features in the WBCD dataset, whereas each row (Subset) means one of the obtained feature subsets from one method. The square (i, j) will be highlighted if the j-th feature is included in the i-th feature subset. The color becomes darker as the subset size (the number of features included) increases, i.e., the feature subsets are ranked based on their sizes. Meanwhile, the figure also gives the training accuracy (on the right side) of the learning algorithm using each feature subset and the frequency (in the upper side) with which a feature has been selected across the obtained subsets.

As shown in Fig. 6, when the size of a feature subset is larger than 1, there will be two or more common features included in the obtained feature subsets. For example, four features F_2, F_3, F_{22}, and F_{23}, are common features in feature subsets S_3-S_6, and features F_{22} and F_{23} are common features in feature subsets with size greater than or equal to 2. This shows that the four features F_2, F_3, F_{22}, and F_{23} are more likely to be strongly relevant features. According to the description from UCI Machine Learning Repository [18], Features F_2 and F_{22} are the mean and the largest value of *texture*–standard deviation of gray-scale values of cell nuclei in all the sampled images, and features F_3 and F_{23} are the mean and the largest value of *perimeter* of cell nuclei. This suggests that the *texture* and *perimeter* of cell nuclei are two essential factors in the diagnosis of Breast Cancer in Wisconsin. On the other hand, the irrelevant features e.g., F_{14} and F_{25}, are easy to identify by their low frequency, which helps us categorize the remaining variables as weakly relevant features.

Another interesting point is that although two different feature subsets $\{F_2, F_3, F_{21}, F_{22}, F_{23}, F_{27}\}$ and $\{F_1, F_2, F_3, F_{21}, F_{22}, F_{23}\}$ in Fig. 6, achieve the same highest training accuracy (94.968%), they may have different feature collection costs. More specifically, F_1 is the mean radius of cell nuclei, while F_{27} is the largest value of *concavity*–severity of concave portions of the contour. F_1 is easier to collec than F_{27}. Therefore, if both feature subsets are provided to a user, the user is more likely to choose the second feature subset since it has lower feature collection cost.

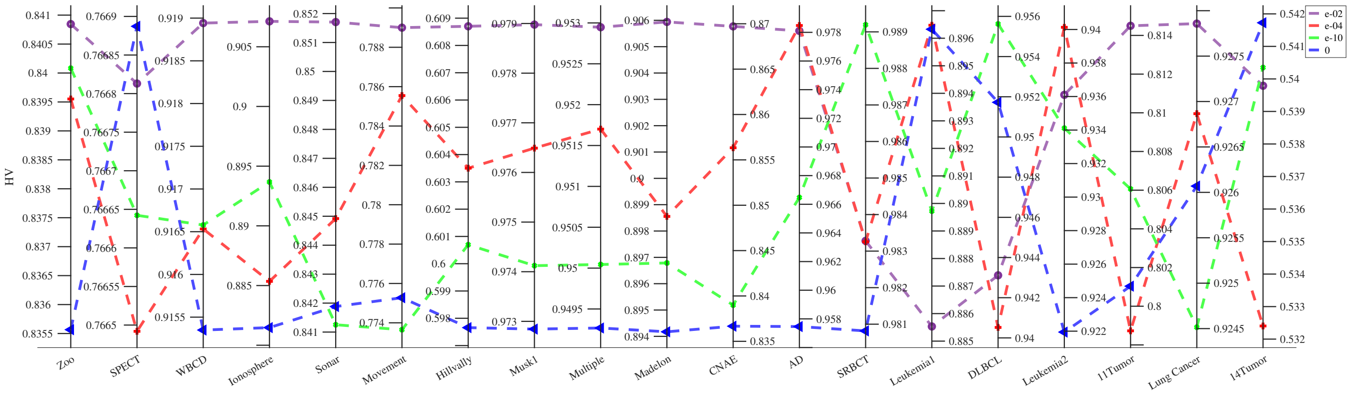

Fig. 7. The average HV results of NMDE with different ψ values.

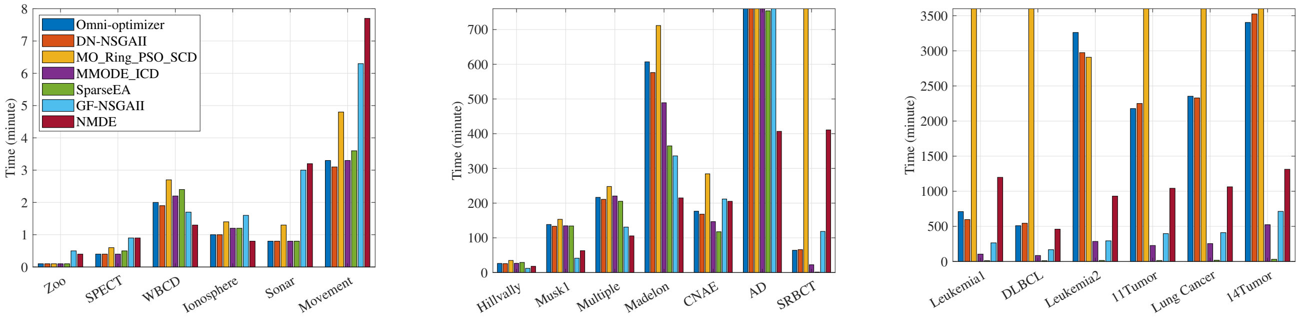

Fig. 8. Running time consumed by the seven algorithms on the 19 datasets (unit: minute).

E. Parameter (ψ) Analysis

This work transforms finding multiple optimal feature subsets into searching for ψ-quasi equal feature subsets, thus it is necessary to examine the effect of ψ on the performance. The average HV results of ψ at $e-02$, $e-04$, $e-10$, and 0 are reported in Fig. 7.

As mentioned above, the larger the HV, the better the algorithm. As shown in Fig. 7, $\psi = e-02$ achieves the highest HV values on 12 out of the 19 datasets. On eight datasets (Zoo, SPECT, WBCD, Ionosphere, SRBCT, DLBCL, 11Tumor, and 14Tumor), ψ at $e-10$ obtains better HV results than ψ at $e-4$, while this trend is opposite on the remaining 11 dataset. In terms of ψ at 0, it gets the lowest average HV values on most of the used datasets. However, ψ at 0 achieves the largest HV values on SPECT and 14Tumor datasets. In addition, the difference of the HV values among the four methods do not exceed 0.2.

The results show that NMDE is not very sensitive to the values of ψ. Meanwhile, $\psi = e-10$ means that the difference of classification accuracy between two feature subsets is close to 0. When ψ is larger, e.g., $e-02$ or $e-04$, the performance slightly increases in general, but the value, $e-02$, may relax the classification accuracy a lot. When ψ is equal to 0, the HV performance slightly decreases. Therefore, a slightly larger value of ψ than 0 is recommended. Based on Fig. 7, $\psi = e-04$ is a good starting point.

F. Efficiency of NMDE

All experiments are carried out using the Mahuika High-Performance Computing (HPC) cluster of the New Zealand eScience Infrastructure (NeSI). For an intuitive analysis, the average computational times of the seven algorithms for the 19 datasets in minutes are divided into three figures, as shown in Fig. 8.

From Fig. 8, the slowest algorithm among the seven methods is MO_Ring_PSO_SCD, since MO_Ring_PSO_SCD consumes the longest time on 13 out of the 19 datasets. For instance, the average training time (over 3000 minutes) of MO_Ring_PSO_SCD on the DLBCL dataset is more than six times longer than that of NMDE (less than 500 minutes). In addition, Omni-optimizer, DN-NSGAII, and MO_Ring_PSO_SCD consume over twice longer than other methods on the last four high-dimensional datasets. This is because the size of feature subsets obtained from Omni-optimizer, DN-NSGAII, and MO_Ring_PSO_SCD is much larger (shown in Fig. 5 and Fig. S.2 in the Online Supplementary Materials). For a wrapper-based feature selection method, classification performance evaluation is the most time-consuming step because of the involvement of the classification process. This is also the reason that the average training time of SparseEA is the lowest among the seven methods on the last seven high-dimensional datasets.

NMDE spends the highest training time only on the Sonar (3.2 minutes) and Movement (7.7 minutes) datasets. On the large datasets which represent harder feature selection tasks, although NMDE finishes the evolutionary training process in a longer time than MMODE_ICD, SparseEA, and GF-NSGAII, the overall training time is less than 24 hours, which is acceptable in most cases, particularly for non-real-time scenarios that this paper is focused on.

VI. Conclusions and Future Work

The goal of this paper was to design a new method to find multiple optimal solutions in feature selection tasks of varying difficulty. The goal has been successfully achieved by transforming the task of finding multiple optimal feature subsets into searching for ψ-quasi equal feature subsets. The proposed NMDE algorithm introduced a new mutation operator, a tailored environmental selection strategy, and a subset repairing mechanism. NMDE was examined and compared with six multi-objective feature selection algorithms on 19 datasets with a different numbers of features ranging from 16 to over $15,000$. The results showed that NMDE outperformed all the six compared methods in terms of both the HV and IGD performance metrics. More importantly, NMDE successfully found different feature subsets with similar or the same classification performance. Moreover, NMDE selected a much smaller number of features while achieving higher accuracy on most of the 19 datasets. Further analyses indicated that the proposed mutation operator can speed up the convergence as well as produce promising feature subsets, the tailored environmental selection strategy can preserve ψ-quasi equal feature subsets, and the proposed subset repairing mechanism can decrease the redundant rate of feature subsets. Combining them into NMDE leads to the final best performance.

In the future, we will apply ensemble learning techniques to NMDE, and study the classification performance of combining different feature subsets with the same classification accuracy.

References

[1] H. B. Nguyen, B. Xue, P. Andreae, and M. Zhang, "Particle swarm optimisation with genetic operators for feature selection," in *IEEE Congr. Evol. Comput.*, 2017, pp. 286–293.

[2] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," *IEEE Trans. Evol. Comput.*, vol. 20, no. 4, pp. 606–626, 2015.

[3] B. Tran, B. Xue, and M. Zhang, "Variable-length particle swarm optimization for feature selection on high-dimensional classification," *IEEE Trans. Evol. Comput.*, vol. 23, no. 3, pp. 473–487, 2018.

[4] G. Karakaya, S. Galelli, S. D. Ahipaşaoğlu, and R. Taormina, "Identifying (quasi) equally informative subsets in feature selection problems for classification: a max-relevance min-redundancy approach," *IEEE Trans. Cybern.*, vol. 46, no. 6, pp. 1424–1437, 2015.

[5] P. Wang, B. Xue, J. Liang, and M. Zhang, "Improved crowding distance in multi-objective optimization for feature selection in classification." Springer, 2021, pp. 489–505.

[6] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journ. Mach. Learn. Research*, vol. 5, pp. 1205–1224, 2004.

[7] H. Liu and H. Motoda, *Feature extraction, construction and selection: a data mining perspective*. Springer Science & Business Media, 1998, vol. 453.

[8] X. Feng, S. Wang, Q. Liu, H. Li, J. Liu, C. Xu, W. Yang, Y. Shu, W. Zheng, B. Yu *et al.*, "Selecting multiple biomarker subsets with similarly effective binary classification performances," *Journ. Vis. Exper.*, no. 140, 2018.

[9] J. Liu, C. Xu, W. Yang, Y. Shu, W. Zheng, and F. Zhou, "Multiple similarly effective solutions exist for biomedical feature selection and classification problems," *Scient. Reports*, vol. 7, no. 1, pp. 1–10, 2017.

[10] C. Yue, J. Liang, B. Qu, K. Yu, and H. Song, "Multimodal multiobjective optimization in feature selection," in *IEEE Congr. Evol. Comput.*, 2019, pp. 302–309.

[11] S. Kamyab and M. Eftekhari, "Feature selection using multimodal optimization techniques," *Neurocomputing*, vol. 171, pp. 586–597, 2016.

[12] C. Qiu and X. Zuo, "Barebones particle swarm optimization with a neighborhood search strategy for feature selection," in *Intern. Conf. Bio-Insp. Comp. Theor. Appl.* Springer, 2018, pp. 42–54.

[13] H. Xia, J. Zhuang, and D. Yu, "Combining crowding estimation in objective and decision space with multiple selection and search strategies for multi-objective evolutionary optimization," *IEEE Trans. Cybern.*, vol. 44, no. 3, pp. 378–393, 2013.

[14] S. Ronald, "Finding multiple solutions with an evolutionary algorithm," in *IEEE Intern. Conf. Evol. Comput.*, vol. 2, 1995, pp. 641–646.

[15] X. Li, M. G. Epitropakis, K. Deb, and A. Engelbrecht, "Seeking multiple solutions: an updated survey on niching methods and their applications," *IEEE Trans. Evol. Comput.*, vol. 21, no. 4, pp. 518–538, 2016.

[16] C. Yue, B. Qu, and J. Liang, "A multiobjective particle swarm optimizer using ring topology for solving multimodal multiobjective problems," *IEEE Trans. Evol. Comput.*, vol. 22, no. 5, pp. 805–817, 2017.

[17] H. Zhao, Z. Zhan, Y. Lin, X. Chen, X. Luo, J. Zhang, S. Kwong, and J. Zhang, "Local binary pattern-based adaptive differential evolution for multimodal optimization problems," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3343–3357, 2019.

[18] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[19] I. Kononenko, "Estimating attributes: Analysis and extensions of relief," in *Europ. Conf. Mach. Learn.* Springer, 1994, pp. 171–182.

[20] H. Xu, B. Xue, and M. Zhang, "A duplication analysis based evolutionary algorithm for bi-objective feature selection," *IEEE Trans. Evol. Comput.*, vol. 25, no. 2, pp. 205–218, 2020.

[21] R. Tanabe and H. Ishibuchi, "A review of evolutionary multimodal multiobjective optimization," *IEEE Trans. Evol. Comput.*, vol. 24, no. 1, pp. 193–200, 2019.

[22] F. Zaman, S. M. Elsayed, T. Ray, and R. A. Sarkerr, "Evolutionary algorithms for finding nash equilibria in electricity markets," *IEEE Trans. Evol. Comput.*, vol. 22, no. 4, pp. 536–549, 2017.

[23] D. Whitley, "A genetic algorithm tutorial," *Statist. Comp.*, vol. 4, no. 2, pp. 65–85, 1994.

[24] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Intern. Conf. Neural Networks*, vol. 4. IEEE, 1995, pp. 1942–1948.

[25] R. Storn and K. Price, "Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces," *Journ. Global Optimiz.*, vol. 11, no. 4, pp. 341–359, 1997.

[26] C. Yue, P. Suganthan, J. Liang, B. Qu, K. Yu, Y. Zhu, and L. Yan, "Differential evolution using improved crowding distance for multimodal multiobjective optimization," *Swarm Evol. Comput.*, vol. 62, p. 100849, 2021.

[27] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Machine learning proceedings*. Elsevier, 1992, pp. 249–256.

[28] H. Almuallim and T. G. Dietterich, "Learning boolean concepts in the presence of many irrelevant features," *Artif. Intell.*, vol. 69, no. 1-2, pp. 279–305, 1994.

[29] M. A. Hall and L. A. Smith, "Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper." in *FLAIRS Conf.*, vol. 1999, 1999, pp. 235–239.

[30] L. Yu and H. Liu, "Feature selection for high-dimensional data: a fast correlation-based filter solution," in *Intern. Conf. Mach. Learn.*, 2003, pp. 856–863.

[31] S. J. Reeves and Z. Zhe, "Sequential algorithms for observation selection," *IEEE Trans. Sign. Process.*, vol. 47, no. 1, pp. 123–132, 1999.

[32] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Patt. Recogn. Lett.*, vol. 15, no. 11, pp. 1119–1125, 1994.

[33] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Patt. Analys. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, 2005.

[34] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: a data perspective," *ACM Comp. Surveys*, vol. 50, no. 6, pp. 1–45, 2017.

[35] X. Li, Y. Wang, and R. Ruiz, "A survey on sparse learning models for feature selection," *IEEE Trans. Cybern.*, 2020.

[36] G. C. Cawley, N. L. Talbot, and M. Girolami, "Sparse multinomial logistic regression via bayesian l1 regularisation," *Advances Neural Inform. Process. Syst.*, vol. 19, p. 209, 2007.

[37] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint $l_{2,1}$-norms minimization," *Advances Neural Inform. Process. Syst.*, vol. 23, pp. 1813–1821, 2010.

[38] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, 2002.

[39] B. Xue, W. Fu, and M. Zhang, "Multi-objective feature selection in classification: A differential evolution approach," in *Asia-Pacif. Conf. Simulated Evol. Learn.* Springer, 2014, pp. 516–528.

[40] E. Hancer, "A differential evolution approach for simultaneous clustering and feature selection," in *Intern. Conf. Artif. Intell. Data Process.* IEEE, 2018, pp. 1–7.

[41] Y. Zhang, D. Gong, X. Gao, T. Tian, and X. Sun, "Binary differential evolution with self-learning for multi-objective feature selection," *Inform. Sciences*, vol. 507, pp. 67–85, 2020.

[42] A. A. Bidgoli, H. Ebrahimpour-Komleh, and S. Rahnamayan, "A novel multi-objective binary differential evolution algorithm for multi-label feature selection," in *IEEE Congr. Evol. Comput.*, 2019, pp. 1588–1595.

[43] X. Song, Y. Zhang, D. Gong, and X. Gao, "A fast hybrid feature selection based on correlation-guided clustering and particle swarm optimization for high-dimensional data," *IEEE Trans. Cybern.*, DOI: 10.1109/TCYB.2021.3061152.

[44] B. H. Nguyen, B. Xue, P. Andreae, H. Ishibuchi, and M. Zhang, "Multiple reference points-based decomposition for multiobjective feature selection in classification: static and dynamic mechanisms," *IEEE Trans. Evol. Comput.*, vol. 24, no. 1, pp. 170–184, 2019.

[45] F. Han, W. Chen, Q. Ling, and H. Han, "Multi-objective particle swarm optimization with adaptive strategies for feature selection," *Swarm Evol. Comput.*, vol. 62, p. 100847, 2021.

[46] P. Wang, B. Xue, M. Zhang, and J. Liang, "A grid-dominance based multi-objective algorithm for feature selection in classification," in *IEEE Congr. Evol. Comput.*, 2021, pp. 2053–2060.

[47] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimization for feature selection in classification: a multi-objective approach," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1656–1671, 2012.

[48] Y. Hu, Y. Zhang, and D. Gong, "Multiobjective particle swarm optimization for feature selection with fuzzy cost," *IEEE Trans. Cybern.*, vol. 51, no. 2, pp. 874–888, 2020.

[49] Y. Tian, X. Zhang, C. Wang, and Y. Jin, "An evolutionary algorithm for large-scale sparse multiobjective optimization problems," *IEEE Trans. Evol. Comput.*, vol. 24, no. 2, pp. 380–393, 2019.

[50] H. Xu, B. Xue, and M. Zhang, "Segmented initialization and offspring modification in evolutionary algorithms for bi-objective feature selection," in *Genet. Evol. Comput. Conf.*, 2020, pp. 444–452.

[51] K. A. De Jong, "Analysis of the behavior of a class of genetic adaptive systems," University of Michigan, Tech. Rep., 1975.

[52] D. E. Goldberg, J. Richardson *et al.*, "Genetic algorithms with sharing for multimodal function optimization," in *Intern. Conf. Genet. Alg.* Hillsdale, NJ: Lawrence Erlbaum, 1987, pp. 41–49.

[53] J. Li, M. E. Balazs, G. T. Parks, and P. J. Clarkson, "A species conserving genetic algorithm for multimodal function optimization," *Evol. Comput.*, vol. 10, no. 3, pp. 207–234, 2002.

[54] K. Jha and S. Saha, "Incorporation of multimodal multiobjective optimization in designing a filter based feature selection technique," *Applied Soft Comp.*, p. 106823, 2020.

[55] K. Deb and S. Tiwari, "Omni-optimizer: a generic evolutionary algorithm for single and multi-objective optimization," *Europ. Journ. Operat. Research*, vol. 185, no. 3, pp. 1062–1087, 2008.

[56] J. Liang, C. Yue, and B. Qu, "Multimodal multi-objective optimization: a preliminary study," in *IEEE Congr. Evol. Comput.*, 2016, pp. 2454–2461.

[57] Z. Zhao, L. Wang, and H. Liu, "Efficient spectral feature selection with minimum redundancy," in *AAAI Conf. Artif. Intell.*, vol. 24, no. 1, 2010.

[58] L. Xu, Q. Zhou, A. Huang, W. Ouyang, and E. Chen, "Feature selection with integrated relevance and redundancy optimization," in *IEEE Intern. Conf. Data Mining*. IEEE, 2015, pp. 1063–1068.

[59] X. Xu, X. Wu, F. Wei, W. Zhong, and F. Nie, "A general framework for feature selection under orthogonal regression with global redundancy minimization," *IEEE Trans. Knowl. Data Engin.*, 2021.