

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. Each solution in the set 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 choosing 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. Second, a newly developed environmental selection mechanism allows equal informative feature subsets to be stored by relaxing the Pareto-dominance relationship. Lastly, the proposed subset repairing mechanism can generate better feature subsets and further remove the redundant features. The proposed method is compared against seven 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 (test instances). The training phase of the classification algorithm aims to discover the relationship between instances' properties (features) and their class labels.

Manuscript received November 26, 2021; revised March 2, 2022; accepted April 5, 2022. This work was supported in part by the Marsden Fund of New Zealand Government under Contracts VUW1913 and VUW1914, National Natural Science Foundation of China (NSFC) under Grant 61922072 and 61876169, the Science for Technological Innovation Challenge (SfTI) fund under grant E3603/2903, MBIE Data Science SSIF Fund under the contract RTVU1914, MBIE Research Program under MBIE C11X2001 and FPR Contract 39186, China Scholarship Council (CSC) and Victoria University Joint Scholarship. (*Corresponding author: Jing Liang*)

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.

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 $2^n - 1$ possible feature subsets. 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 without affecting the classification performance. Irrelevance indicates a feature is not necessary at all. Weak relevance indicates that a feature may become useful when it is combined with other features, although the feature 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 may result 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 therefore 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/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 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 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 to 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 and 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. Furthermore, 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 address the above limitations. The major contributions are as follows:

- 1) 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. The aim is 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 a lower classification error rate.
- 2) NMDE transforms the task of searching for multiple optimal feature subsets into searching for ψ -quasi equal feature subsets. This is 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.
- 3) 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 the 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

A. Conventional Feature Selection Methods

Feature selection methods can be roughly classified into the wrapper, filter, and embedded methods. Wrapper methods evaluate the goodness of a feature subset utilizing a specific learning algorithm, 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.

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 are 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, 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 for the fitness evaluations. 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 the center of PF [44], [45]. To cover this issue, different dominance relationships including grid-dominance [46], ε -dominance [47], and fuzzy dominance [48] have been applied to feature selection. 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). However, some parameters in the proposed MMODE_ICD algorithm, e.g., the selection ratio of solutions from different fronts, are not easy to set when applying MMODE_ICD to feature selection.

Another issue 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. Xu *et al.* [49] designed an offspring modification mechanism to improve the population diversity by using the features' selected frequency. 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 up one of them may not be a good choice, and the users are likely to choose according to their domain knowledge [9]. To search for multiple optimal feature subsets, Wang *et al.* [5] employed the 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 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 using the EMO technique with extreme learning machine to find multiple optimal feature subsets.

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 [50], sharing-based [51], and speciation-based [52] methods, are to incorporate

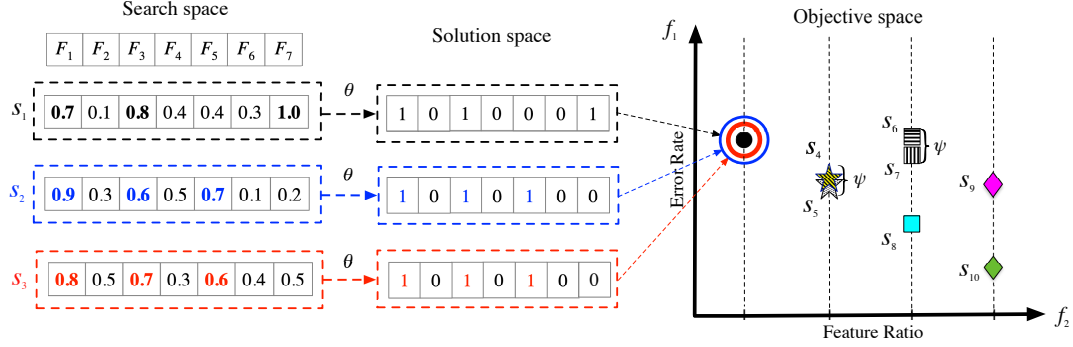


Fig. 1. Situations of ψ -quasi equal feature subsets in the search space, the solution space, and the objective space. Suppose a dataset has seven features, F_1 to F_7 , and S_1 to S_{10} represent ten feature subsets on this feature selection task. According to an individual's representation in Section 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 .

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. The proposed special crowding distance mechanism considers the crowding estimations in both the search space and the objective space. Kanchan et al. [53] further improved the method in [10] and developed a filter-based multimodal feature selection method. However, both methods from [10], [53] need a large population to introduce niching behavior.

Although several niching-based feature selection methods have been proposed, the feature subsets obtained from these methods such as in [10] and [53] still include some 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 details of the proposed NMDE method. The notations used are summarized in Table I.

A. Representation and Objective Functions

The representation of an individual (shown in Fig. 1) in NMDE is a vector with real-valued encoding (between 0 and 1). The dimensionality of each vector equals 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 (position entry) is not less than 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.

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

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

TABLE I: Notations Used in This Paper

\mathcal{N}_i	Niche of the current individual \vec{x}_i
\mathcal{S}_i	Individuals of \mathcal{N}_i at the first front
\vec{x}_{nal}	Individual in \mathcal{S}_i with the lowest classification error rate
Qe	A set of multiple groups of ψ -quasi equal feature subsets
TS_1-TS_t	t ψ -quasi equal feature subsets in one group of Qe
r_{min}	Lowest front number of solutions TS_1-TS_t
$freq$	Selected frequency of each feature in $\{TS_1-TS_t\}$
P	Current population
O	Offspring
PO	Combination of P and O
P_u	A set storing all unique solutions during evolution
Sf	A newly produced feature subset
ln	Upper bound of the number of ψ -quasi equal feature subsets in each group
$\mathcal{W}_1-\mathcal{W}_k$	k Pareto fronts
C_i	Final crowding distance of individual \vec{x}_i
C_i^o	Crowding distance of individual \vec{x}_i in the objective space
C_i^s	Crowding distance of individual \vec{x}_i in the solution space

selected features ($\#Selected\ Features$) over the total number of original features ($\#Original\ Features$).

B. Equivalence of Feature Subsets

Definition 1: For a given measure of the classification error rate (f_1), the ratio of the number of features selected in a feature subset (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 the same f_1 and f_2 values, e.g., S_1, S_2, S_3 in Fig. 1.

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.

C. Niching Behavior

In NMDE, each individual (\vec{x}_i) has its 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]. In 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*, 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).

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

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 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. Duplicated feature subsets, e.g., S_2 and S_3 in Fig. 1, can also lead to the same objective values, thus the duplicated feature subsets should be removed first. After that, the presence of the ψ -quasi equal feature subsets includes three situations. Fig. 1 shows *Situation 1* where some common features are included in the ψ -quasi equal feature subsets (e.g., 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

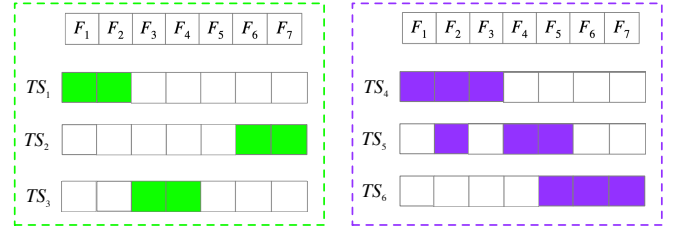


Fig. 2. Another two situations of ψ -quasi equal feature subsets.

Algorithm 1: Subset Repairing Operator

Input: P , O , P_u , and ψ
Output: PO , P_u , and Qe

```

1 begin
2   Get  $PO = P \cup O$  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  $i = 1, \dots, |Qe|$  do
6     Get  $t$  equally feature subsets,  $TS_1 \dots TS_t$ , 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       $PO = PO \cup Sf$  and  $P_u = P_u \cup Sf$ ,
15    end
16  end
17 end

```

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 on the right side of Fig. 2 where some solutions have common features.

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). 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, \dots, TS_t (Line 6 in Alg. 1). Next, the selected frequency of each feature is computed based on $\{TS_1, TS_2, \dots, TS_t\}$, termed $freq$ (Line 7 in Alg. 1).

For *Situations 1* and *3* (Line 8 in Alg. 1), a new feature subset Sf is created by setting its j -th variable Sf_j with the following formula:

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

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

Eq. (3) shows that one feature will be selected from Sf only when the selected frequency value of the feature is larger than

Algorithm 2: ψ -quasi Equal Subsets Grouping

Input: PO , PO_fit , and ψ
Output: Qe

```

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

```

$1/t$. Therefore, Sf will select fewer number of features than each feature subset in $\{TS_1, TS_2, \dots, TS_t\}$. If the classification error rate of Sf is not larger than that of a solution in $\{TS_1, TS_2, \dots, TS_t\}$, that means redundant features are included in $\{TS_1, TS_2, \dots, TS_t\}$. During the environmental selection process, Sf will be preferred. Otherwise, all the solutions in $\{TS_1, TS_2, \dots, TS_t\}$ are preferred.

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

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

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 $\{TS_1, TS_2, \dots, TS_t\}$, Sf may have a lower error rate since the combined features may provide different complementary information to the class labels.

To save computational resources, all feature subsets produced and their objective values will be preserved into a set 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 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.

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_4 from the second front will be dragged into the first front. Since S_4 is ψ -quasi equal subset to S_5 , S_5 is at the first front. Likewise, feature subsets S_6 and S_7 will also sit at the same front. 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 keeps all of them during evolution, that will reduce

Algorithm 3: Environmental Selection

Input: PO , Qe , and ψ
Output: P

```

1 begin
2   Rank  $PO$  into  $k$  fronts ( $\mathcal{W}_1, \mathcal{W}_2, \dots, \mathcal{W}_k$ ) via non-dominated
   sorting, and set  $P = \emptyset$ ,
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: $\mathcal{W}_1, \mathcal{W}_2, \dots, \mathcal{W}_k$ and Qe
Output: $\mathcal{W}_1, \mathcal{W}_2, \dots, \mathcal{W}_k$

```

1 begin
2   for  $i = 1, \dots, |Qe|$  do
3     Get  $TS_1, TS_2, \dots, TS_t$  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

```

the diversity of solutions in the objective space. For example, suppose that five solutions need to be chosen from Fig. 3 to form the population for the next generation, the algorithm may select solution S_1 - S_5 . However, the other solutions (S_8 - S_{10}) with a lower classification error rate 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.

$$ln = |PO| // \text{unique_number}(f_2(\mathcal{W}_1)) \quad (5)$$

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 $FR(f_2)$ values. Using Eq. (5), we can get $ln = 10 // 4 = 2$. Therefore, NMDE will pick up two solutions with a larger crowding distance value from S_1 - S_7 . Then, the selected two solutions with S_8 - S_{10} form a new population for the next generation. The calculations of the crowding distance are introduced as follows.

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 descending order based on their crowding distances. Meanwhile, the individuals in the first $m - 1$ fronts are put in a temporary set \mathcal{T} .

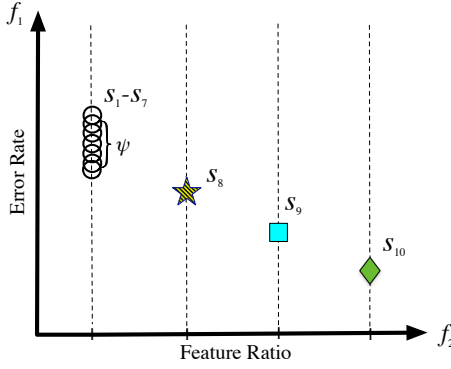


Fig. 3. An example of the distribution of the objective values of 10 solutions. Based on the **Definition 1**, S_1 to S_7 are ψ -quasi equal feature subsets.

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

$$C_i = \begin{cases} \max\{C_i^o, C_i^s\} & \text{if } C_i^o > C_{\text{avg}}^o \text{ or } C_i^s > C_{\text{avg}}^s \\ \min\{C_i^o, C_i^s\} & \text{otherwise} \end{cases} \quad (6)$$

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.

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

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, $\text{obj} = 2$.

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., $C_6^o = C_7^o$.

Algorithm 5: NMDE

Input: NP , Max_FEs , and ψ
Output: P

```

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

```

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).

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

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. 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

Alg. 5 gives the complete NMDE algorithm. 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.

IV. EXPERIMENT DESIGN

A. Benchmark Techniques

Seven EMO methods are chosen for comparisons with NMDE, i.e., Omni-optimizer [54], MO_Ring_PSO_SCD [16], MMODE_ICD [26], SparseEA [55], SM-MOEA [56], DAEA [20], and GF-NSGAI [46]. Among them, Omni-optimizer, MO_Ring_PSO_SCD, and MMODE_ICD are three multimodal multi-objective optimization algorithms, which are used for finding multiple optimal solutions of a multimodal problem. SparseEA, SM-MOEA, DAEA, and GF-NSGAI are four EMO-based feature selection methods. SparseEA and DAEA

TABLE II: 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	Hillvalley	100	2	1,212
8	Musk1	166	2	2,031
9	Multiple (pix)	240	10	2,000
10	Madelon	500	2	2,600
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

are implemented using PlatEMO [57], and the remaining 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 II. The datasets 13-19 are from the biomedical domain, which include thousands of features but tens of instances (SRBCT and DLBCL) and up to 10,000 features and a few hundreds of instances (11Tumor and Lung Cancer). 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. The specific parameters of the compared algorithms are from the corresponding literatures [16], [20], [26], [46], [54]–[56]. For NMDE, F in Eq. (2) is set to 0.5, and the crossover rate (CR) is set to 0.5 [26]. The threshold θ is set to 0.6, as suggested in [3], [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 NP for all the 19 datasets.

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. The K value in KNN is set to five so as to balance the accuracy and efficiency.

For a feature selection task, considering the cross-validation, the minimal interval between the adjacent values of the classification accuracy is within the range of $[0, 1/T]$ where T means the number of training instances in a dataset. The lowest and the largest number of training instances are $72 \times 0.7 \approx 50$ and $3,279 \times 0.7 \approx 2,295$ on Leukemia2 and AD, respectively. Therefore, the value of $1/T$ on the two datasets is $1/50 = 2e - 02$ and $1/2,295 \approx 4.4e - 04$, respectively. To avoid over-loosening the constraint on the classification

accuracy, it is reasonable to make ψ in the range $[0, 5/T)$. Section V.F compares and analyses the performance of NMDE with $\psi = 1/T$, $\psi = 2/T$, $\psi = 3/T$, $\psi = 4/T$, and 0. This paper mainly discusses the experimental results at $\psi = 1/T$.

C. Performance Indicators

Six performance indicators are used to compare different feature selection methods. They are 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) [58]–[60], which is defined as:

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

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 III-IV show the average HV and IGD results of the eight algorithms on the 19 datasets. The reference point of HV is set to (1, 1). For calculating IGD , all the output feature subsets from the eight algorithms are collected and merged into one set. Then, all feature subsets in the set are ranked by the non-dominated sorting. 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 III-IV, 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 significant difference from NMDE, respectively. The Wilcoxon test with a significance level of 0.05 is used. Furthermore, the Freidman test is employed to give the relative performance ranking among the eight algorithms.

A. Overall Analysis

The results show that NMDE achieves the best among the eight methods in terms of the test HV and IGD results.

As shown in Table III among the 133 comparisons with the other seven algorithms, there are only six losses of NMDE on HV . One significant loss of NMDE happens against SparseEA on the AD test set, and one loss happens against SM-MOEA on the SPECT test set. The other four losses happen against DAEA on the SPECT, Musk1, AD, and Lung Cancer datasets, respectively, but NMDE still takes the second-best on the Musk1 and Lung Cancer datasets among the eight algorithms. The IGD performance shown in Table IV reveals

TABLE III: Average *HV* results on test sets (the larger the better).

Dataset	Omni-optimizer	MO_Ring_PSO_SCD	MMODE_ICD	SparseEA	SM-MOEA	DAEA	GF-NSGAI	NMDE
Zoo	7.877e-01 ±4.539e-02 ↓	8.476e-01 ±1.126e-02 ^o	8.350e-01 ±1.180e-02 ↓	8.369e-01 ±8.390e-03↓	8.241e-01 ±3.553e-02↓	8.377e-01 ±7.330e-03↓	8.276e-01 ±1.259e-02↓	8.507e-01 ±1.843e-02
SPECT	6.946e-01 ±4.730e-02 ↓	7.593e-01 ±1.320e-02 ^o	7.650e-01 ±0.000e+00 ^o	7.668e-01 ±4.700e-03 ^o	7.921e-01 ±2.325e-02↑	7.935e-01 ±3.168e-02↑	7.666e-01 ±3.700e-03 ^o	7.665e-01 ±8.440e-03
WBCD	8.030e-01 ±3.621e-02 ↓	9.134e-01 ±8.930e-03 ^o	9.071e-01 ±1.116e-02 ↓	9.099e-01 ±1.130e-02↓	9.043e-01 ±9.340e-03↓	9.141e-01 ±9.370e-03 ^o	9.117e-01 ±4.240e-03↓	9.155e-01 ±7.100e-03
Ionosphere	7.461e-01 ±5.002e-02 ↓	8.933e-01 ±1.906e-02↓	8.880e-01 ±2.699e-02 ↓	8.934e-01 ±3.164e-02 ^o	8.698e-01 ±3.811e-02↓	8.958e-01 ±2.170e-02 ^o	8.857e-01 ±1.800e-02↓	9.030e-01 ±1.229e-02
Sonar	7.127e-01 ±3.189e-02 ↓	8.091e-01 ±3.008e-02↓	8.468e-01 ±2.945e-02 ^o	8.473e-01 ±3.054e-02 ^o	7.982e-01 ±4.316e-02↓	8.445e-01 ±2.485e-02 ^o	8.241e-01 ±3.710e-02↓	8.542e-01 ±3.325e-02
Movement	6.442e-01 ±1.995e-02 ↓	7.109e-01 ±1.303e-02↓	7.536e-01 ±2.603e-02 ↓	7.758e-01 ±1.659e-02 ^o	7.263e-01 ±3.029e-02↓	7.784e-01 ±1.233e-02 ^o	7.438e-01 ±2.172e-02↓	7.783e-01 ±1.537e-02
Hillvalley	4.755e-01 ±1.529e-02 ↓	5.262e-01 ±1.189e-02↓	5.815e-01 ±1.427e-02 ↓	5.928e-01 ±7.470e-03 ^o	5.778e-01 ±8.130e-03↓	5.887e-01 ±6.000e-03↓	5.932e-01 ±1.054e-02 ^o	5.955e-01 ±6.660e-03
Musk1	7.686e-01 ±1.995e-02 ↓	8.132e-01 ±1.025e-02↓	9.105e-01 ±1.931e-02 ↓	9.709e-01 ±3.100e-03↓	9.609e-01 ±5.290e-03↓	9.785e-01 ±2.060e-03↑	9.579e-01 ±7.020e-03↓	9.760e-01 ±2.930e-03
Multiple	7.526e-01 ±1.444e-02 ↓	7.686e-01 ±9.450e-03↓	8.299e-01 ±1.533e-02 ↓	9.514e-01 ±2.030e-03 ^o	9.401e-01 ±7.190e-03↓	9.474e-01 ±4.010e-03↓	8.925e-01 ±1.445e-02↓	9.509e-01 ±2.830e-03
Madelon	5.893e-01 ±1.473e-02 ↓	6.157e-01 ±6.450e-03↓	7.925e-01 ±2.424e-02 ↓	8.997e-01 ±5.920e-03 ^o	8.916e-01 ±1.220e-02↓	8.914e-01 ±4.980e-03↓	8.896e-01 ±3.960e-03↓	9.009e-01 ±4.640e-03
CNAE	5.931e-01 ±1.593e-02 ↓	6.167e-01 ±9.200e-03↓	7.070e-01 ±1.382e-02 ↓	8.656e-01 ±8.730e-03↓	8.413e-01 ±1.926e-02↓	7.923e-01 ±6.240e-03↓	8.290e-01 ±1.012e-02↓	8.759e-01 ±1.228e-02
AD	6.807e-01 ±9.760e-03 ↓	6.897e-01 ±3.440e-03↓	7.476e-01 ±1.245e-02 ↓	9.795e-01 ±2.270e-03↑	9.711e-01 ±2.540e-03↓	9.792e-01 ±1.320e-03↑	8.988e-01 ±1.237e-02↓	9.781e-01 ±1.780e-03
SRBCT	6.623e-01 ±1.911e-02 ↓	6.922e-01 ±1.052e-02↓	7.357e-01 ±1.672e-02 ↓	9.180e-01 ±3.677e-02↓	9.073e-01 ±4.754e-02↓	9.592e-01 ±3.419e-02↓	8.942e-01 ±1.258e-02↓	9.781e-01 ±2.434e-02
Leukemia1	5.640e-01 ±3.157e-02 ↓	5.949e-01 ±2.614e-02↓	6.195e-01 ±3.371e-02 ↓	8.786e-01 ±6.120e-02↓	8.263e-01 ±6.863e-02↓	8.607e-01 ±7.271e-02↓	6.998e-01 ±5.920e-02↓	9.171e-01 ±5.117e-02
DLBCL	6.552e-01 ±1.664e-02 ↓	6.830e-01 ±7.510e-03↓	7.070e-01 ±8.950e-03 ↓	9.618e-01 ±4.576e-02 ^o	9.090e-01 ±4.947e-02↓	9.279e-01 ±5.386e-02 ^o	7.896e-01 ±1.543e-02↓	9.467e-01 ±5.317e-02
Leukemia2	5.970e-01 ±1.622e-02 ↓	6.234e-01 ±1.240e-02↓	6.369e-01 ±1.343e-02 ↓	9.297e-01 ±3.869e-02 ^o	8.952e-01 ±4.492e-02↓	9.379e-01 ±4.619e-02 ^o	7.197e-01 ±1.944e-02↓	9.342e-01 ±3.449e-02
11Tumor	5.265e-01 ±1.956e-02 ↓	5.586e-01 ±1.384e-02↓	5.688e-01 ±1.578e-02 ↓	7.761e-01 ±3.879e-02↓	7.098e-01 ±6.461e-02↓	8.007e-01 ±3.989e-02↓	6.357e-01 ±2.446e-02↓	8.237e-01 ±3.757e-02
Lung Cancer	6.234e-01 ±9.030e-03 ↓	6.415e-01 ±8.600e-03↓	6.660e-01 ±9.900e-03 ↓	9.033e-01 ±3.089e-02↓	8.995e-01 ±3.482e-02↓	9.392e-01 ±1.881e-02↑	7.201e-01 ±1.132e-02↓	9.272e-01 ±1.794e-02
14Tumor	3.474e-01 ±1.168e-02 ↓	3.635e-01 ±1.200e-02↓	3.757e-01 ±9.160e-03 ↓	4.987e-01 ±1.448e-02↓	4.833e-01 ±4.889e-02↓	5.378e-01 ±3.188e-02 ^o	3.992e-01 ±1.148e-02↓	5.462e-01 ±2.558e-02
Rank	8.00 (0/0/19)	6.16 (0/3/16)	5.53 (0/2/17)	2.61 (1/9/9)	3.77 (1/0/18)	2.41 (4/7/8)	4.81 (0/2/17)	1.44

a similar pattern to *HV*. There are seven losses out of the 133 comparisons of NMDE on *IGD*. Specifically, NMDE loses to SparseEA on three datasets, i.e., the SPECT, Madelon, and CNAE datasets. In addition, MO_Ring_PSO_SCD, MMODE_ICD, SM-MOEA, and DAEA achieve better *IGD* performance than NMDE only on the SPECT 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, Madelon, SRBCT, and 11Tumor datasets, are chosen since their PF distributions from the eight 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 seven compared algorithms on both the training and test sets. In addition, it is obvious although some feature subsets obtained by the eight algorithms have the same training and/or testing classification accuracy, the feature subsets from NMDE include fewer features than that from Omni-optimizer, MO_Ring_PSO_SCD,

and MMODE_ICD. For example, on the SRBCT dataset, although different feature subsets from the eight algorithms can achieve both the highest training and testing classification accuracy (nearly 100%), the feature subsets from NMDE only contain less than 4% of the original features. However, this number for Omni-optimizer, MO_Ring_PSO_SCD, and MMODE_ICD is over 30%. For GF-NSGAI, this number is higher than 10%. Although the number of selected features among SparseEA, SM-MOEA, DAEA, and NMDE is close, the feature subsets from NMDE have better distribution in the objective space.

In summary, NMDE wins 210, draws 43 and losses 13 out of the 266 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 and converge better than those of the seven compared EMO algorithms.

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 two feature subsets in Fig. 6, will be kept in

TABLE IV: Average *IGD* results on test sets (the smaller the better).

Dataset	Omni-optimizer	MO_Ring_PSO_SCD	MMODE_ICD	SparseEA	SM-MOEA	DAEA	GF-NSGAI	NMDE
Zoo	1.069e-01 ±2.832e-02 ↓	5.126e-02 ±1.137e-02 ^o	6.143e-02 ±1.182e-02 ↓	6.401e-02 ±9.090e-03 ↓	7.801e-02 ±2.796e-02 ↓	6.414e-02 ±9.480e-03 ↓	6.858e-02 ±8.800e-03 ↓	4.843e-02 ±7.900e-03
SPECT	6.850e-02 ±3.070e-02 ↓	3.060e-02 ±1.210e-02 ↑	2.540e-02 ±7.400e-03 ↑	2.660e-02 ±1.110e-02 ↑	3.245e-02 ±3.183e-02 ↑	3.521e-02 ±2.413e-02 ↑	4.420e-02 ±1.570e-02 ^o	4.010e-02 ±1.070e-02
WBCD	1.094e-01 ±3.215e-02 ↓	2.180e-02 ±6.810e-03 ↓	2.289e-02 ±6.920e-03 ↓	1.980e-02 ±7.040e-03 ↓	2.487e-02 ±6.220e-03 ↓	1.909e-02 ±6.320e-03 ↓	1.964e-02 ±4.050e-03 ↓	1.306e-02 ±4.960e-03
Ionosphere	1.293e-01 ±3.776e-02 ↓	3.297e-02 ±1.190e-02 ↓	3.836e-02 ±1.627e-02 ↓	3.050e-02 ±2.077e-02 ^o	5.384e-02 ±2.723e-02 ↓	3.133e-02 ±1.606e-02 ^o	3.872e-02 ±9.800e-03 ↓	2.543e-02 ±8.190e-03
Sonar	1.608e-01 ±2.730e-02 ↓	8.884e-02 ±1.444e-02 ↓	6.991e-02 ±1.785e-02 ↓	6.991e-02 ±1.820e-02 ^o	1.005e-01 ±3.489e-02 ↓	7.438e-02 ±1.900e-02 ↓	7.606e-02 ±2.455e-02 ↓	5.937e-02 ±2.174e-02
Movement	2.249e-01 ±2.132e-02 ↓	1.590e-01 ±1.358e-02 ↓	1.032e-01 ±3.979e-02 ↓	5.204e-02 ±2.510e-03 ↓	7.724e-02 ±1.817e-02 ↓	4.979e-02 ±8.750e-03 ↓	7.117e-02 ±1.899e-02 ↓	4.425e-02 ±7.990e-03
Hillvally	1.673e-01 ±2.292e-02 ↓	9.784e-02 ±1.200e-02 ↓	3.856e-02 ±1.536e-02 ↓	3.083e-02 ±5.040e-03 ^o	4.956e-02 ±5.850e-03 ↓	3.289e-02 ±4.300e-03 ↓	3.436e-02 ±4.650e-03 ↓	2.880e-02 ±5.030e-03
Musk1	1.812e-01 ±2.185e-02 ↓	1.387e-01 ±1.004e-02 ↓	5.018e-02 ±1.455e-02 ↓	1.466e-02 ±2.540e-03 ↓	3.023e-02 ±5.270e-03 ↓	1.183e-02 ±1.720e-03 ^o	2.746e-02 ±4.140e-03 ↓	1.203e-02 ±2.550e-03
Multiple	2.423e-01 ±1.353e-02 ↓	2.274e-01 ±9.490e-03 ↓	1.683e-01 ±1.425e-02 ↓	2.072e-02 ±2.910e-03 ^o	2.881e-02 ±3.400e-03 ↓	2.848e-02 ±1.122e-02 ↓	9.699e-02 ±1.493e-02 ↓	2.001e-02 ±3.230e-03
Madelon	3.067e-01 ±1.217e-02 ↓	2.898e-01 ±5.550e-03 ↓	1.526e-01 ±2.066e-02 ↓	1.632e-02 ±3.640e-03 ↑	2.098e-02 ±6.370e-03 ^o	9.735e-02 ±3.206e-02 ↓	1.221e-01 ±5.430e-03 ↓	2.501e-02 ±1.878e-02
CNAE	3.050e-01 ±1.047e-02 ↓	3.009e-01 ±7.250e-03 ↓	2.441e-01 ±1.243e-02 ↓	2.691e-02 ±4.990e-03 ↑	4.374e-02 ±1.000e-02 ↓	1.848e-01 ±7.850e-03 ↓	1.634e-01 ±6.190e-03 ↓	3.446e-02 ±9.240e-03
AD	3.004e-01 ±1.009e-02 ↓	2.924e-01 ±3.530e-03 ↓	2.306e-01 ±1.319e-02 ↓	2.587e-02 ±4.760e-03 ^o	3.372e-02 ±1.690e-03 ↓	2.657e-02 ±1.510e-03 ^o	8.876e-02 ±9.780e-03 ↓	2.690e-02 ±3.270e-03
SRBCT	3.680e-01 ±8.940e-03 ↓	3.546e-01 ±5.330e-03 ↓	3.223e-01 ±1.453e-02 ↓	4.883e-02 ±1.827e-02 ^o	5.549e-02 ±2.170e-02 ↓	3.228e-02 ±1.221e-02 ^o	1.956e-01 ±9.860e-03 ↓	3.952e-02 ±2.390e-02
Leukemia1	3.952e-01 ±1.466e-02 ↓	3.694e-01 ±1.074e-02 ↓	3.412e-01 ±1.561e-02 ↓	8.593e-02 ±4.525e-02 ↓	1.323e-01 ±5.188e-02 ↓	1.010e-01 ±5.380e-02 ↓	2.344e-01 ±3.175e-02 ↓	4.957e-02 ±3.850e-02
DLBCL	3.705e-01 ±4.160e-03 ↓	3.493e-01 ±2.370e-03 ↓	3.227e-01 ±7.450e-03 ↓	3.132e-02 ±3.754e-02 ^o	7.293e-02 ±4.051e-02 ↓	6.044e-02 ±4.380e-02 ^o	2.266e-01 ±4.870e-03 ↓	4.204e-02 ±4.280e-02
Leukemia2	3.819e-01 ±6.798e-03 ↓	3.173e-01 ±7.329e-03 ↓	3.442e-01 ±6.352e-03 ↓	4.194e-02 ±3.096e-02 ^o	7.477e-02 ±4.155e-02 ↓	3.948e-02 ±3.595e-02 ^o	2.552e-01 ±7.473e-03 ↓	4.265e-02 ±2.9449e-02
11Tumor	3.938e-01 ±3.680e-03 ↓	3.641e-01 ±3.550e-03 ↓	3.604e-01 ±5.980e-03 ↓	8.274e-02 ±2.621e-02 ↓	1.245e-01 ±4.862e-02 ↓	5.147e-02 ±2.063e-02 ^o	3.066e-01 ±6.390e-03 ↓	5.129e-02 ±1.476e-02
Lung Cancer	3.006e-01 ±3.426e-03 ↓	2.834e-01 ±5.633e-03 ↓	2.647e-01 ±4.843e-03 ↓	1.095e-01 ±1.351e-02 ↓	1.180e-01 ±3.223e-02 ↓	9.940e-02 ±5.980e-03 ^o	2.249e-01 ±2.443e-03 ↓	9.767e-02 ±9.906e-03
14Tumor	3.894e-01 ±2.670e-03 ↓	3.572e-01 ±6.450e-03 ↓	3.572e-01 ±6.450e-03 ↓	5.211e-02 ±1.747e-02 ↓	5.814e-02 ±2.614e-02 ↓	2.428e-02 ±1.213e-02 ^o	3.068e-01 ±3.910e-03 ↓	2.500e-02 ±8.000e-03
Rank	8.00 (0/0/19)	6.28 (1/1/17)	5.50 (1/0/18)	2.34 (3/8/8)	4.81 (1/1/17)	2.75 (1/9/9)	4.75 (0/1/18)	1.65

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 size 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, Hillvally, Multiple, Madelon, Leukemia1, 11Tumor, and 14Tumor datasets are chosen as representative to show the performance in Fig. 5. More results on the remaining 11 datasets are shown in Table S.V-S.VI in the Online Supplementary Materials due to the page limit.

In Fig. 5, NMDE has the largest *Num* value apart from the 11Tumor dataset. The highest improvement can be seen on the Leukemia1 dataset, NMDE can find over 55 different feature subsets with the same training accuracy, while this number on the other methods is less than 5. More importantly, those feature subsets from NMDE on the Leukemia1 dataset include less than 100 features but with higher test accuracy than many compared methods. Furthermore, NMDE achieves the highest accuracy on the WBCD, Hillvally, and 14Tumor datasets with fewer features included than that from Omni-optimizer, MO_Ring_PSO_SCD, and MMODE_ICD. The same patterns can also be seen in the Sonar, Movement and Multiple datasets

from Table S.V-S.VI in the Online Supplementary Materials. On the SRBCT and Leukemia1 datasets, NMDE can achieve the second-best of accuracy among the eight methods. On the 11Tumor dataset, the superiority of NMDE on the accuracy slightly decreases, but NMDE still has a significant advantage on *Size* results over GF-NSGAI.

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 are shown in Tables S.VII-S.VIII 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. The five methods are compared with a new DE-based EMO method (named MODE-NM) with the proposed mutation operator.

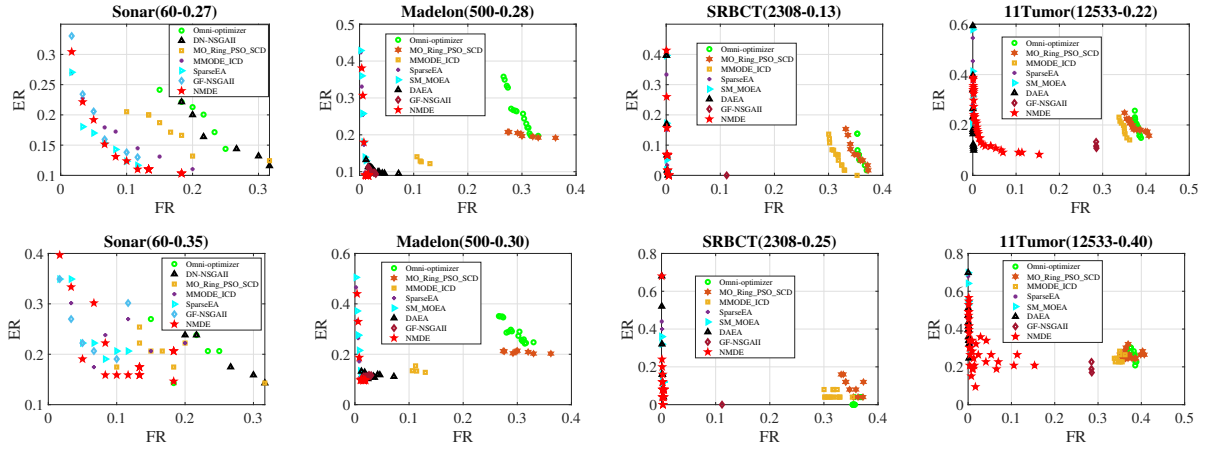
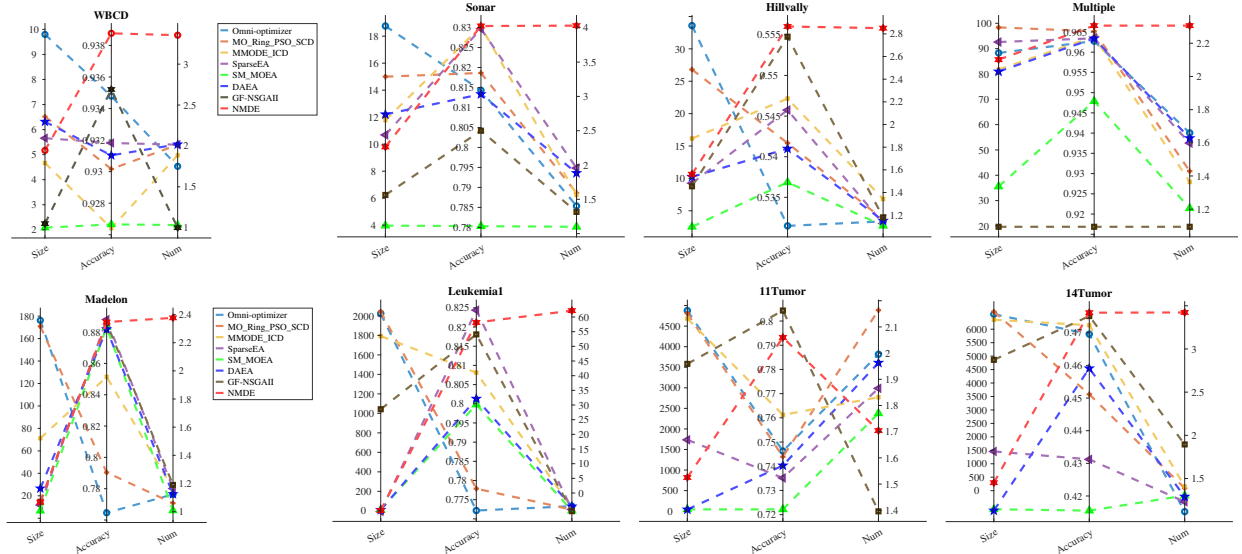


Fig. 4. Obtained PFs of eight algorithms on the training and the test sets (upper row for training results and bottom row for test results).

Fig. 5. Average test results of the feature subsets with the lowest training error rate from the eight 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*.

Only the mutation operator differs among the six methods. The *HV*, *IGD* results, and average *HV* plots with generations of the six methods are shown in Table S.VII, Table S.VIII, and Fig. S.3 in the Online Supplementary Materials, respectively.

In Fig. S.3, MODE-NM achieves the fast convergence and the largest *HV* values than MODE-rand1, MODE-best1, MODE-current-rand1, MODE-rand2, and MODE-best2 on almost all the training sets. More importantly, MODE-NM achieves significantly better rankings than the other five methods on both test *HV* and *IGD* results. The results show that the proposed mutation operator accelerates the convergence of the algorithm and obtains feature subsets with better feature selection 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.VII-S.VIII show that the proposed environmental selection strategy can help NMDE to achieve better *HV* and *IGD* results. In Section V.B, the results show that the proposed environmental selection strategy can

TABLE V: Average *RED* results between NMDE-N (NMDE without the proposed subset repairing mechanism) and NMDE.

	NMDE-N	NMDE		NMDE-N	NMDE
Zoo	1.646e-01 ±2.001e-02	1.651e-01 ±3.085e-02	CNAE	4.926e-01 ±6.559e-04	4.827e-01 ±4.151e-04
SPECT	4.170e-02 ±2.336e-02	6.675e-02 ±2.948e-02	AD	4.924e-01 ±6.157e-04	4.284e-01 ±1.749e-02
WBCD	1.998e-02 ±1.718e-03	3.930e-03 ±7.844e-04	SRBCT	3.692e-01 ±9.622e-03	2.737e-01 ±4.333e-02
Ionosphere	2.354e-01 ±5.395e-02	2.149e-01 ±4.327e-02	Leukemia1	2.774e-01 ±1.470e-02	1.822e-01 ±6.992e-02
Sonar	2.917e-02 ±3.066e-03	2.689e-02 ±3.501e-03	DLBCL	4.697e-01 ±1.517e-02	1.278e-01 ±8.971e-02
Movement	4.687e-03 ±6.512e-04	3.665e-03 ±6.158e-04	Leukemia2	3.439e-01 ±9.420e-03	1.276e-01 ±3.538e-02
Hillvally	1.152e-05 ±1.751e-06	1.155e-05 ±2.135e-06	11Tumor	4.349e-01 ±2.027e-02	3.696e-01 ±9.754e-03
Musk1	5.485e-01 ±2.203e-02	5.542e-01 ±2.590e-02	Lung Cancer	3.142e-01 ±1.933e-02	2.528e-01 ±1.621e-02
Multiple	1.057e-01 ±6.550e-03	1.000e-01 ±4.265e-03	14Tumor	3.637e-01 ±7.526e-03	3.495e-01 ±4.154e-03
Madelon	4.513e-04 ±1.123e-04	1.714e-05 ±4.514e-05	Sum	1 / 6 / 12	N/A

help NMDE to find more ψ -quasi equal feature subsets.

3) *Subset Repairing Strategy*: NMDE is formed using all three major components. The results in Tables S.VII-S.VIII show that the proposed subset repairing mechanism by modifying the ψ -quasi equal feature subsets during evolution can help NMDE to achieve higher *HV* and/or lower *IGD* values

in most of the 19 datasets. Table V gives the *RED* results between NMDE and NMDE-N. In Table V, the feature subsets obtained from NMDE show a significantly lower redundancy rate than that from NMDE-N on 12 out of the 19 datasets. On 6 datasets including the Zoo, Ionosphere, Sonar, Hillvally, Musk1, and Multiple datasets, the *RED* values between NMDE and NMDE-N have no significant difference. The results suggest that the proposed subset repairing scheme helps NMDE to obtain feature subsets with a 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. 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_{11} , 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_{10} and F_{16} , 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 collect 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.

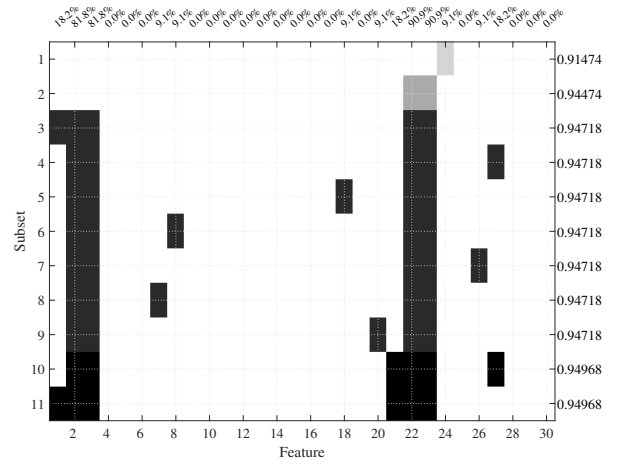


Fig. 6. Frequency matrix from NMDE on the WBCD dataset. ‘Feature’ means one of the 30 original features in the WBCD dataset, and ‘Subset’ means one of the obtained feature subsets from NMDE. The training accuracy using each feature subset is shown on the right side, and the selected frequency of each feature is shown on the upper side.

E. Further Analysis on Different Learning Algorithms

Similar or the same feature selection performance can also be influenced by different training algorithms, different initialization methods, and/or randomness in a learning algorithm. To show the existence of multiple optimal feature subsets regardless of the classification algorithm used, Section S.VII in the Online Supplementary Materials performs further experiments by using more different classification algorithms. The results show that although different learning algorithms are used, NMDE still can find that different feature subsets with the same size achieve the same classification accuracy.

F. Parameter (ψ) Analysis

The average *HV* results of ψ at 0, $1/T$, $2/T$, $3/T$, and $4/T$ are reported in Fig. 7. As mentioned above, the larger the *HV*, the better the algorithm. In Fig. 7, $\psi = 4/T$ achieves the highest *HV* values on 8 out of the 19 datasets. On 12 datasets, ψ at $3/T$ obtains better *HV* results than ψ at $2/T$, while this trend is opposite on the remaining 7 dataset (Zoo, Sonar, Hillvally, CNAE, AD, Leukemia1, and Lung Cancer). In terms of ψ at 0, it gets the lowest average *HV* values on most of the used datasets. However, ψ at 0 achieves larger *HV* values on SPECT, Hillvally, SRBCT, and DLBCL datasets than ψ at $1/T$. In addition, the difference of the *HV* values among the five methods does not exceed 0.2.

The results show that NMDE is not very sensitive to the values of ψ . When ψ is larger, e.g., $3/T$ or $4/T$, the performance generally slightly increases, but the value might over relax in terms of the classification accuracy. When ψ is equal to 0, the *HV* performance slightly decreases. Therefore, a slightly larger value than 0 is recommended for ψ . Based on Fig. 7, $\psi = 1/T$ is a good starting point.

G. Efficiency of NMDE

All experiments are carried out using the Mahuika High-Performance Computing (HPC) cluster of the New Zealand eScience Infrastructure (NeSI) [61]. For an intuitive analysis, the average computational times of the eight algorithms for

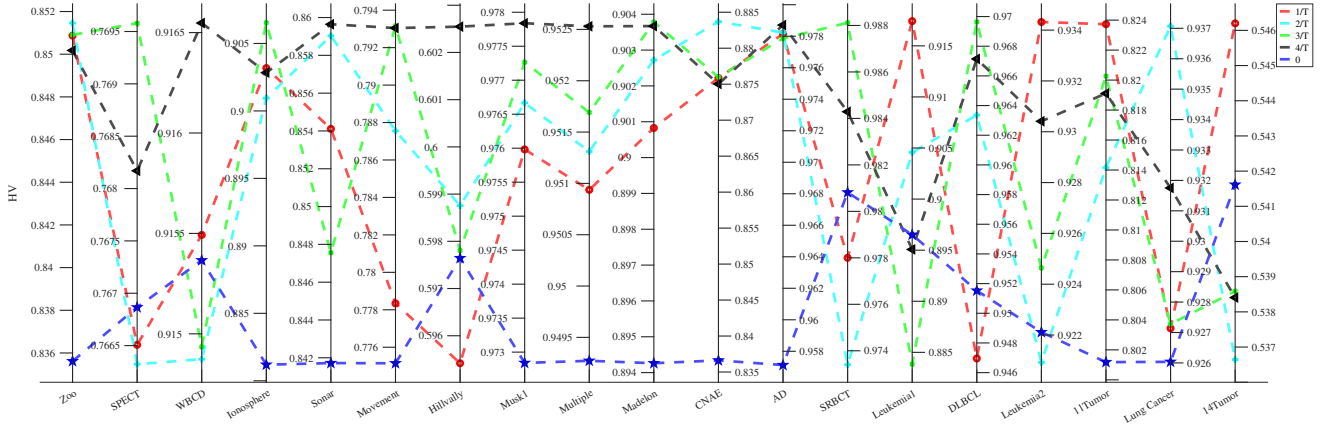
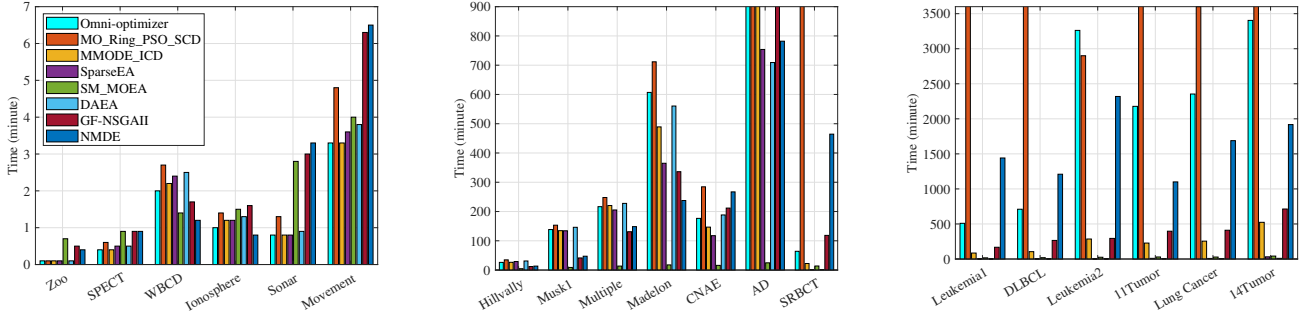
Fig. 7. Average HV results of NMDE with different ψ values.

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

the 19 datasets in minutes are divided into three figures, as shown in Fig. 8.

From Fig. 8, the slowest algorithm among the eight methods is MO_Ring_PSO_SCD, since MO_Ring_PSO_SCD consumes the longest time on 13 out of the 19 datasets. In addition, Omni-optimizer and MO_Ring_PSO_SCD consume the longest time on the last four high-dimensional datasets. This is because the size of feature subsets obtained from Omni-optimizer and MO_Ring_PSO_SCD is much larger. For a wrapper-based feature selection method, classification performance evaluation is the most time-consuming step because of the involvement of the classification process.

NMDE spends the highest training time only on the Sonar (3.3 minutes) and Movement (6.5 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, SM-MOEA, DAEA, and GF-NSGAI, the overall training time is less than 48 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. 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 seven multi-objective feature selection algorithms on 19 datasets with different numbers of features ranging from 16 to over 15,000. The results showed that NMDE outperformed all the seven compared methods in terms of both the HV and IGD performance metrics. Also, NMDE successfully found different feature subsets with similar or the same classification performance. Furthermore, 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 and 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 extend NMDE to solve other real-world problems in the domains with multiple optimal solutions requirements. Also, 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," in *Intern. Conf. Appl. Evol. Comput.* 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. Andreea, 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] 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.
- [50] K. A. De Jong, "Analysis of the behavior of a class of genetic adaptive systems," University of Michigan, Tech. Rep., 1975.
- [51] 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.
- [52] 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.
- [53] K. Jha and S. Saha, "Incorporation of multimodal multiobjective optimization in designing a filter based feature selection technique," *Applied Soft Comp.*, p. 106823, 2020.

- [54] 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.
- [55] 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.
- [56] F. Cheng, F. Chu, Y. Xu, and L. Zhang, "A steering-matrix-based multi-objective evolutionary algorithm for high-dimensional feature selection," *IEEE Trans. Cybern.*, DOI: 10.1109/TCYB.2021.3053944.
- [57] Y. Tian, R. Cheng, X. Zhang, and Y. Jin, "Platemo: A matlab platform for evolutionary multi-objective optimization [educational forum]," *IEEE Computat. Intell. Magazine*, vol. 12, no. 4, pp. 73–87, 2017.
- [58] Z. Zhao, L. Wang, and H. Liu, "Efficient spectral feature selection with minimum redundancy," in *AAAI Conf. Artif. Intell.*, vol. 24, no. 1, 2010.
- [59] 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.
- [60] 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.
- [61] A. Pletzer, W. Hayek, C. Scott, B. Corrie, and G. Rae, "How NeSI helps users run better and faster on new zealand's supercomputing platforms," in *IEEE Intern. Conf. e-Science*, 2017, pp. 465–466.



Peng Wang (Graduate Student Member, IEEE) received the B.E. degree from Zhengzhou University of Aeronautics, China, in 2016 and the M.S. degree from Zhengzhou University, China, in 2019. He is now pursuing a Ph.D. degree in Victoria University of Wellington, New Zealand.

His current research interests include evolutionary computation, multi-objective optimization, and feature selection.



Bing Xue (Senior Member, IEEE) received the B.Sc. degree from the Henan University of Economics and Law, Zhengzhou, China, in 2007, the M.Sc. degree in management from Shenzhen University, Shenzhen, China, in 2010, and the Ph.D. degree in computer science in 2014 at VUW, New Zealand.

She is currently a Professor of Artificial Intelligence, and Deputy Head of School in the School of Engineering and Computer Science at VUW. She has over 300 papers published in fully refereed

international journals and conferences and her research focuses mainly on evolutionary computation, machine learning, classification, symbolic regression, feature selection, evolving deep NNs, image analysis, transfer learning, multi-objective machine learning.

Dr Xue is currently the Chair of IEEE CIS Evolutionary Computation Technical Committee, Chair of IEEE CIS Task Force on Evolutionary Deep Learning and Applications, and Editor of IEEE CIS Newsletter. She has also served as associate editor of several international journals, such as IEEE Computational Intelligence Magazine, IEEE Transactions on Evolutionary Computation and ACM Transactions on Evolutionary Learning and Optimization.



Jing Liang (Senior Member, IEEE) received the B.E. degree from Harbin Institute of Technology, China, in 2003 and the Ph.D. degree from the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, in 2009.

She is currently a Professor with the School of Electrical Engineering, Zhengzhou University, Zhengzhou, China. Her main research interests are evolutionary computation, swarm intelligence, multiobjective optimization, and neural network. She currently serves as an Associate Editor for the IEEE

Transactions on Evolutionary Computation, and the Swarm and Evolutionary Computation.



Mengjie Zhang (Fellow, IEEE) received the B.E. and M.E. degrees from Artificial Intelligence Research Center, Agricultural University of Hebei, Hebei, China, and the Ph.D. degree in computer science from RMIT University, Melbourne, VIC, Australia, in 1989, 1992, and 2000, respectively.

He is currently Professor of Computer Science, Head of the Evolutionary Computation Research Group, and the Associate Dean (Research and Innovation) in the Faculty of Engineering, VUW. His current research interests include evolutionary computation,

particularly genetic programming and particle swarm optimization with application areas of image analysis, multi-objective optimization, feature selection and reduction, job shop scheduling, and transfer learning. He has published over 700 papers in refereed international journals and conferences.

Prof. Zhang is a Fellow of the Royal Society of New Zealand, a Fellow of IEEE, an IEEE CIS Distinguished Lecturer, and have been a Panel Member of the Marsden Fund (New Zealand Government Funding). He was the chair of the IEEE CIS Intelligent Systems and Applications Technical Committee, the chair of the IEEE CIS Emergent Technologies Technical Committee and the Evolutionary Computation Technical Committee, and a member of the IEEE CIS Award Committee.