SUPERVISED FEATURE SELECTION VIA MULTIOBJECTIVE PROGRAMMING AND ITS APPLICATION IN THE MEDICAL FIELD

PHAM THI KHANH[1], PHAM THI HOAI [1,*]

[1] *Faculty of Mathematics and Informatics, Hanoi University of Science and Technology, 1 Dai Co Viet Road, Hanoi, Vietnam.*

ABSTRACT. In this study, we model the supervised feature selection problem using a novel approach: convex bi-objective optimization. Traditional methods have addressed this problem by maximizing relevance to class labels and minimizing redundancy among features. Recently, Wang et al. [30] formulated this problem as a single-objective convex optimization, yielding only a unique solution. Unlike that, we approach this problem by preserving the natural multi-objective essence of the supervised feature selection problem, enabling a broader exploration of the objective space and providing a set of optimal solutions rather than a single result. To solve the obtained model, we utilize two state-of-the-art strategies: an exact method and a heuristic method. Additionally, we have enhanced the exact method using a scaling technique, which accelerates processing speed and expands the Pareto front. In parallel, the heuristic method ensures that the Pareto solutions achieve extensive coverage and distribution. The effectiveness of our proposed method is confirmed through standard medical datasets, demonstrating superiority over existing techniques. Notably, in the context of skin cancer screening, the method optimized the feature set to less than half of its original size, thereby significantly enhancing classification accuracy for the task.

Keywords. supervised feature selection, multiobjective programming, constrained multiobjective optimization problems, convex multiobjective programming, objective normalization, scaling technique

1. INTRODUCTION

To improve the performance of machine learning models on high-dimensional data, dimensionality reduction is an essential preprocessing step. This technique helps reduce computational complexity, storage requirements, and training time. Currently, dimensionality reduction is mainly done through two approaches: feature extraction and feature selection [10, 12]. While feature extraction transforms the original data into a new space where variables lose their physical meaning, feature selection keeps the original attributes. This characteristic makes feature selection better for interpretability, as it allows users to see a direct link between the original features and the model's predictions [24, 13] .

By removing noisy and redundant features, feature selection helps prevent the loss of important information, which improves learning accuracy and classification performance. Because of these benefits, this technique is widely used in complex fields such as image processing, data mining, bioinformatics (including healthcare), and natural language processing [30, 1, 17].

Based on how label information is used, feature selection methods are generally divided into three groups: unsupervised, semi-supervised, and supervised [17]. Among these, supervised feature selection directly uses class labels to guide the evaluation and selection process, aiming to find the best feature subset for the model [17, 30]. Traditional supervised feature selection methods are categorized into three main types: filter, wrapper, and embedded [5, 16]. Wrapper methods use the performance of a specific machine learning algorithm as a criterion to find the best features [16]. Although they are usually accurate, they are computationally expensive because of the repeated training process [16, 30]. Embedded methods integrate the selection process directly into the training phase of the model. This approach balances speed and accuracy, but the results depend heavily on the specific algorithm used [29,

*Corresponding author.
E-mail addresses: hoai.phamthi@hust.edu.vn; phamhoai051087@gmail.com; khanh207111@gmail.com
2020 Mathematics Subject Classification: 47N10, 58E7, 46N10

33]. Finally, filter methods rank features based on their relevance to class labels and can process high-dimensional data quickly. However, the effectiveness of filter methods depends on the search strategy, which generally follows two trends: the incremental approach and the holistic approach [30].

The incremental approach selects features one by one based on evaluation scores (such as Information Gain, Fisher Score, mRMR, or CIFE) [10, 20, 2, 18, 30]. This approach is simple, has low computational costs, and effectively removes irrelevant features. However, because it is "greedy," it often gets stuck in local optima; once a feature is selected, it cannot be removed later. Moreover, some incremental methods assume features are independent, so they cannot handle redundancy well and are sensitive to noise [30]. In contrast, the holistic approach considers the interactions between all features at the same time (e.g., QPFS, GRM, AGRM) [21, 19, 28, 30], which helps avoid the local optima problem of incremental methods. However, current holistic methods often use unsupervised measures for redundancy (ignoring class labels), which reduces their effectiveness in classification tasks [30].

To solve these problems, Wang et al.[30] proposed a supervised holistic feature selection method based on neurodynamic optimization. By considering all feature interactions and using label-based mutual information, this algorithm avoids "greedy" mistakes and filters out redundant features more accurately for classification. The authors tested this method on eight benchmark datasets, showing that it converges very fast (in milliseconds) and performs better than common methods like IG, Fisher Score, mRMR and CIFE. Despite these strengths, the method has some weaknesses: it requires data discretization which may lose original information, and it models the problem in a way that only provides one single optimal feature subset. This means it misses the chance to find other useful feature subsets that might also satisfy the model's goals.

Our main contributions: To overcome the limitations mentioned above, we propose a new approach based on multi-objective optimization. Our model handles the trade-off between minimizing redundancy and maximizing relevance at the same time to find better feature subsets. The resulting formulation of supervised feature selection belongs to the class of convex bi-criteria optimization over a compact and convex set. We then solve the obtained problem by using two state-of-the-art strategies: the first one is an exact method called MPG-Explicit proposed by Bello-Cruz et al. [4]; the second one is a heuristic method named NSGA-II [9] that provides an approximation of the Pareto optimal front. Through empirical analysis, we realize that the magnitude disparity between objectives adversely affects the performance of the exact method - MPG-Explicit. This issue leads us to apply the scaling technique in [11] for MPG-Explicit to obtain a new algorithm named Scaled-MPG-E. Throughout numerical experiments for numerous data sets in the medical field, Scaled-MPG-E significantly speeds up computation time as well as expands the Pareto front compared to the original method. Notably, for large-scale instances, we suggest using the filtering technique [22] that provides quality representative subsets from the Pareto optimal solution set to reduce the processing time without compromising solution quality.

The structure of this paper is as follows. Section 2 reviews some basic concepts and results related to the main contributions of the paper. Section 3 introduces our proposed multi-objective feature selection model in detail, a long with our proposed algorithm, Scaled-MPG-E to solve the resulting optimization problem. In Section 4, we present experimental results to evaluate the efficient performance of our new method with other state-of-the-art algorithms for a lot of benchmark data sets in the medical field. Finally, Section 5 provides some conclusions and future works.

2. PRELIMINARIES

2.1. **Description of supervised feature selection (SFS) problems.**

2.1.1. *Problem description.* Considering a multi-class classification problem involving m classes, n samples, and p features. The training set is defined as $\{(x_i, y_i)|i = 1, \ldots, n\}$, where $x_i = (x_{i1}, \ldots, x_{ip})^T$ represents the p-dimensional feature vector of the i-th sample, and the corresponding label is $y_i \in$

$\{1, \ldots, m\}$. Let $Y = (y_1, \ldots, y_n)^T$ denote the general label vector for the n samples. The data matrix $X = (x^{(1)}, \ldots, x^{(p)})$ of size $n \times p$ is composed of columns $x^{(j)} = (x_{1j}, \ldots, x_{nj})^T$, which serve as the j-th predictors, $j = 1, ..., p$.

To normalize the data, the features are centered to obtain $F_j = C_n x^{(j)}$, where $C_n = I_n - ee^T/n$ is the centering matrix [28] (I_n is the identity matrix and e is the unit vector). The set of all transformed features is denoted as $F = \{F_1, \ldots, F_p\}$. The core objective of the SFS is to identify an optimal subset of k features ($k < p$) that exhibits the least mutual correlation while maintaining the highest relevancy to the target class y. According to this criterion, the k selected features represent the most significant informative components of the entire dataset.

2.1.2. *Supervised redundancy measure.* To quantify the relationship between features, Wang et al.[30] utilize the framework of information theory [7] through the concepts of entropy, mutual information, joint mutual information, and conditional mutual information. In particular, let $\hat{X} = (\hat{x_1}, \ldots, \hat{x_n}), \hat{Y} = (\hat{y_1}, \ldots, \hat{y_n})$ and $\hat{Z} = (\hat{z_1}, \ldots, \hat{z_n})$ be three vectors of discrete random variables. The mathematical quantities are specifically defined as follows:

- Entropy measures the uncertainty of a vector $\hat{X}$ [7]:

$$H(\hat{X}) = -\sum_{\hat{x} \in \hat{X}} p(\hat{x}) \log p(\hat{x}), \tag{2.1}$$

  where $p(\hat{x})$ is the probability density function of the random variable $\hat{x}$. $p(\hat{x})$ can be approximately evaluated with the frequency of $\hat{x}$ appearing in $\hat{X}$ [6], i.e., $p(\hat{x}) = \frac{\mathcal{N}(\hat{x})}{n}$, where $\mathcal{N}(\hat{x})$ is the number of occurrences of value $\hat{x}$ out of n total observations.
- Mutual Information (MI) usually measures the amount of information shared by two random vectors $\hat{X}$ and $\hat{Y}$ [7]:

$$I(\hat{X}; \hat{Y}) = \sum_{\hat{x} \in \hat{X}} \sum_{\hat{y} \in \hat{Y}} p(\hat{x}, \hat{y}) \log \frac{p(\hat{x}, \hat{y})}{p(\hat{x}) p(\hat{y})}, \tag{2.2}$$

  where $p(\hat{x}, \hat{y})$ is the joint probability of $\hat{x}$ and $\hat{y}$. The mutual information and the entropy [7] have the following relation:

$$I(\hat{X}; \hat{Y}) = H(\hat{X}) + H(\hat{Y}) - H(\hat{X}, \hat{Y}). \tag{2.3}$$

- Joint Mutual Information (JMI) [31] is defined as:

$$I(\hat{X}, \hat{Y}; \hat{Z}) = \sum_{\hat{x} \in \hat{X}} \sum_{\hat{y} \in \hat{Y}} \sum_{\hat{z} \in \hat{Z}} p(\hat{x}, \hat{y}, \hat{z}) \log \frac{p(\hat{x}, \hat{y}, \hat{z})}{p(\hat{x}, \hat{y}) p(\hat{z})}, \tag{2.4}$$

  where $p(\hat{x}, \hat{y}, \hat{z})$ is the joint probability of $\hat{x}, \hat{y}$ and $\hat{z}$. The joint mutual information $I(\hat{X}, \hat{Y}; \hat{Z})$ is the amount of information shared by features $\hat{X}, \hat{Y}$ and vector $\hat{Z}$.
- Multi-information $I(\hat{X}; \hat{Y}; \hat{Z})$ [32] is used to isolate the common interaction information between two features and the class label:

$$I(\hat{X}; \hat{Y}; \hat{Z}) = I(\hat{X}; \hat{Y}) + I(\hat{Y}; \hat{Z}) - I(\hat{X}, \hat{Y}; \hat{Z}) \tag{2.5}$$

Based on the above quantities, Wang et al.[30] proposed the supervised redundancy measure s_{ij} to evaluate the redundancy between a pair of features F_i, F_j in correlation with the label y:

$$s_{ij} = \max \left\{ 0, \frac{I(F_i; F_j; y)}{H(F_i) + H(F_j)} \right\}, \tag{2.6}$$

where $I(F_i, F_j, y) = I(F_i; y) + I(F_j; y) - I(F_i, F_j; y)$ based on (2.5).

Proposition 2.1. *([30]) For any features F_i and F_j, the supervised redundancy measure s_{ij} satisfies the following properties:*

(1) *Symmetry:* $s_{ij} = s_{ji}$.
(2) *Normality:* $0 \leq s_{ij} \leq 1$.
(3) *Upper extremity:* $s_{ij} = 1$ *if* F_i *and* F_j *are perfectly correlated with label* y.
(4) *Lower extremity:* $s_{ij} = 0$ *if* F_i *and* F_j *are completely uncorrelated with label* y.

This coefficient strictly adheres to the conditions of a normalized measure, ranging within $[0, 1]$, and accurately reflects the dependency between feature pairs in a supervised environment. The redundancy matrix $S = (s_{ij})$ of size $p \times p$ is used as the basis for evaluating the total redundancy of the feature space. Since the redundancy matrix S may not be guaranteed to be positive semi-definite in all cases, Wang et al.[30]. perform an adjustment on S to obtain a positive semi-definite matrix Q for the optimization model:

$$Q = \delta I_p + S \tag{2.7}$$

where I_p is the identity matrix of the order ρ and the adjustment parameter δ satisfies $\delta \geq -\min\{0, \lambda_{\min}(S)\}$, with $\lambda_{\min}(S)$ being the smallest eigenvalue of S [30]. It is easy to see that Q is a positive semi-definite matrix.

2.1.3. *Feature relevancy measure.* In parallel with the redundancy assessment, Wang et al. [30] quantifies the relevance of each feature using the IG (Information Gain) criterion [2]. Essentially, IG corresponds to the mutual information $I(F_i; y)$, denoted as:

$$\rho_{IG}(F_i) = I(F_i; y). \tag{2.8}$$

The vector $\rho = (\rho_1, \ldots, \rho_p)^T$ represents the relevance levels of all p features with respect to the classification problem.

2.1.4. *The solution method for solving SFS proposed by Wang et al. [30].* To identify an optimal feature subset, Wang et al. [30] proposed a method based on a two-layer recurrent neural network, where the first layer relies on a projection neural network (PNN) for solving the fractional programming problem

$$\min_{w \in C} f(w) = \frac{w^T Q\ w}{\rho^T w} \tag{SFS-Wang et al.}$$

where $w = (w_1, \ldots, w_p)^T$ denotes the feature weight vector and $C = \{w \in \mathbb{R}^p \mid \sum_{i=1}^{p} w_i = 1, w \geq 0\}$. The objective function $f(w)$ results from the desire to minimize the redundancy ($w^T Q w$) and maximize relevancy ($\rho^T w$). Higher weights in the optimal solution w^* of (SFS-Wang et al.) indicate a more significant contribution to the optimal subset. Notably, in Wang et al. [30] the model (SFS-Wang et al.) is recognized as a pseudoconvex programming due to the pseudoconvexity of $f(w)$. However, in the recent work of Hoai et al.[15], this problem is proved to be a convex fractional programming and therefore has a global optimal solution with a unique optimal value. After successfully finding a global optimal solution of (SFS-Wang et al.) w^*, it will be an input for the preceding layer *k-Winners-Take-All* (kWTA) neural network proposed by Wang et al.[27]. This transformation yields the binary vector z^*, whose components (0 or 1) represent the inclusion or exclusion of a specific feature in the final optimal subset.

2.2. **Convex multiobjective programming over a closed convex set.** Considering a multi-objective optimization (MO) over a closed convex $C \subseteq \mathbb{R}^p$ as follows

$$\min_{x \in C} (g_1(x), g_2(x), \ldots, g_\ell(x)), \tag{2.9}$$

where $g_i : \mathbb{R}^p \to \mathbb{R}, i = 1, \ldots, \ell$ are ℓ differentiable and convex objective functions. Unlike single-objective optimization, these objectives are inherently conflicting, whereby enhancing one necessitates a trade-off in the performance of others. As a result, there is typically no unique solution that optimizes all objectives at once. Therefore, the notion of Pareto optimality is utilized to define the set of optimal solutions, as follows:

Definition 2.2. A point $x^* \in C$ is a Pareto optimal solution of problem (2.9) if there is no $x \in C$ such that $g_i(x) \leq g_i(x^*)$ for all $i \in \{1, \ldots, \ell\}$ and $g_j(x) < g_j(x^*)$ for at least one $j \in \{1, \ldots, \ell\}$. Furthermore, $x^* \in C$ is a weakly Pareto optimal point if there does not exist any $x \in C$ such that $g_i(x) < g_i(x^*)$ for all $i = 1, \ldots, \ell$. The collection of all Pareto optimal (weakly Pareto optimal) points constitutes the Pareto set (weakly Pareto set). Their image under $(g_1, ..., g_\ell)$ in the objective space is termed the (weakly) Pareto front.

It is worth noting that Problem (2.9) can be recast in the form of unconstrained multi-objective optimization problems

$$\min_{x \in \mathbb{R}^p} (f_1(x), f_2(x), \ldots, f_\ell(x)), \tag{2.10}$$

where

$$f_j(x) = g_j(x) + h_j(x), \quad j = 1, \ldots, \ell \tag{2.11}$$

and $h_j(x)$ is the indicator of C, i.e., $h_j(x) = i_C(x) = \begin{cases} 0, & \text{if } x \in C \\ +\infty, & \text{otherwise} \end{cases}$. Obviously, the concept of (weakly) Pareto optimality now works on its domain $\bigcap_{j=1}^{\ell} \mathrm{dom} f_j = C$.

In the sequel, we revisited the two efficient methods that can be applied to solve Problem (2.9) or (2.10). The first one is the MPG-Explicit algorithm proposed by Bello-Cruz et al. [4] and the second one is NSGA-II proposed by Deb et al. [9].

2.2.1. *The MPG-Explicit algorithm.* To address Problem (2.10) in the more general setting of h_j such as $h_j : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is proper, convex, and continuous function which is not necessarily differentiable ($j = 1, ..., \ell$), the Multiobjective Proximal Gradient is one of the typical methods [25] can be efficent applied. It determines a descent direction by solving a convex subproblem at each iteration. Specifically, a local model $\psi_{x^k}(\cdot) : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is constructed to approximate the simultaneous variation of the objective functions around the current point x^k:

$$\psi_{x^k}(d) := \max_{j=1,\ldots,\ell} \left\{ \nabla g_j(x^k)^\top d + h_j(x^k + d) - h_j(x^k) \right\}. \tag{2.12}$$

To find the descent direction for Problem (2.10) at x^k, one considers the following subproblem:

$$\min_{d \in \mathbb{R}^p} \psi_{x^k}(d) + \frac{1}{2\alpha}\|d\|^2, \quad \alpha > 0. \tag{2.13}$$

This subproblem has a strongly convex objective, and it has a unique optimal solution $\rho_\alpha(x^k)$. Let $\theta_\alpha(x^k)$ denote for the optimal value of the subproblem (2.13), i.e.,

$$\rho_\alpha(x^k) := \arg\min_{d \in \mathbb{R}^p} \psi_{x^k}(d) + \frac{1}{2\alpha}\|d\|^2 \tag{2.14}$$

and

$$\theta_\alpha(x^k) := \psi_{x^k}(\rho_\alpha(x^k)) + \frac{1}{2\alpha}\|\rho_\alpha(x^k)\|^2. \tag{2.15}$$

The relationship of $\rho_\alpha(x), \theta_\alpha(x)$ and the Pareto optimality of x is stated in the following proposition.

Proposition 2.3. *([25]) Based on the properties of the subproblem* (2.13)*, the following statements are equivalent:*

(i) *x is a weakly Pareto optimal solution of the problem* (2.10)*;*
(ii) $\rho_\alpha(x) = 0$*;*
(iii) $\theta_\alpha(x) = 0$.

Recently, Bello-Cruz et al.[4] introduced an efficient proximal gradient with an explicit line search strategy as follows.

Algorithm 1 (MPG-Explicit)[4]

Step 0: Let $x^0 \in \text{dom}(f)$, $\alpha > 0, \gamma \in (0, 2/\alpha)$, $0 < \tau_1 < \tau_2 < 1$, $k := 0$.
Step 1: Compute $d^k := \rho_\alpha(x^k)$ as in (2.14).
Step 2: **If** $d^k = 0$ **then** return x^k.
Step 3: $j_k* \in \underset{j=1,\dots,\ell}{\text{argmax}} \nabla g_j(x^k)^T d^k$, $t_t := 1$.
Step 3.1: **If** $g_{j_k*}(x^k + t_t d^k) \le g_{j_k*}(x^k) + t_t \nabla g_{j_k*}(x^k)^T d^k + t_t \frac{\gamma}{2}\|d^k\|^2$,
then go to **Step 3.2**.
else $t_t := t_{new}(t_t, d^k, x^k, j_{k^*})$ (by using Procedure 1), back to **Step 3.1**.
Step 3.2: **If** $f(x^k + t_t d^k) \le f(x^k)$ **then** $t_k := t_t$ and go to **Step 4**
else taking the first violated index j_k and go to **Step 3.3**.
Step 3.3: Compute $t_t := t_{new}(t_t, d^k, x^k, j_k)$ (by using Procedure 1).
If for $j = 1, \dots, \ell$,

$$g_j(x^k + t_t d^k) \le g_j(x^k) + t_t \nabla g_j(x^k)^T d^k + t_t \frac{\gamma}{2}\|d^k\|^2$$

then define $t_k = t_t$ and go to **Step 4**.
else back to **Step 3.3** with the first violated index j_k.
Step 4: $x^{k+1} := x^k + t_k d^k$, $k := k + 1$ and back to **Step 1**.

Procedure 1 Computation of $t_{new}(t_t, d^k, x^k, j_k)$ [4]

Input: t_t, d^k, x^k, j_k
Define $\phi(t) := g_{j_k}(x^k + t d^k)$
if $\phi_t'(0) = \nabla g_{j_k}(x^k)^T d^k < 0$ **then** $t_q := -\dfrac{\phi'(0) t_t^2}{2[\phi(t_t) - \phi(0) - \phi'(0) t_t]}$
  if $t_q \in [\tau_1 t_t, \tau_2 t_t]$ **then** $t_{new} := t_q$
  else $t_{new} := t_t/2$
  end if
  $t_{new} := t_t/2$
end if
Output: t_{new}

2.2.2. *NSGA-II method.* NSGA-II is an advanced iteration of the original nondominated sorting genetic algorithm (NSGA), proposed by Deb et al.[9] in 2000. It was specifically designed to overcome the drawbacks of its predecessor, such as high computational complexity ($O(MN^3)$), the absence of an elitism mechanism [23], and a heavy reliance on manual parameter tuning to maintain population diversity. By

incorporating a fast non-dominated sorting procedure and a crowding distance technique, NSGA-II efficiently identifies the Pareto optimal set while ensuring a uniform distribution across the search space. These enhancements reduce the overall computational complexity to $O(MN^2)$ [26], where M represents the number of objectives and N denotes the population size.

3. Our proposed method

3.1. **Reformulation of SFS as a convex bi-criteria optimization problem.** In contrast to the method of Wang et al.[30], which frames the maximum relevance and minimum redundancy objective as a fractional programming problem (SFS-Wang et al.), this paper formulates it as a convex bi-criteria optimization problem as follows

$$\begin{aligned} \min_{w} \quad & \begin{cases} g_1(w) = w^T Q w \\ g_2(w) = -\rho^T w \end{cases} \\ \text{subject to} \quad & \sum_{i=1}^{p} w_i = 1, w \geq 0. \end{aligned} \tag{SFS-MO}$$

We have the following remarkable proposition showing the relation of an optimal solution given by (SFS-Wang et al.) and the Pareto set obtained by (SFS-MO).

Proposition 3.1. *If w^* is an optimal solution of problem* (SFS-Wang et al.)*, then it is a Pareto optimal solution of problem* (SFS-MO).

Proof. Suppose that there exists $\bar{w} \in C = \{w \in \mathbb{R}^p \mid \sum_{i=1}^{p} w_i = 1, w \geq 0\}$ such that $\bar{w}^T Q \bar{w} < w^{*T} Q w^*$ and $-\rho^T \bar{w} \leq -\rho^T w^* < 0$; or $\bar{w} \in C$ such that $\bar{w}^T Q \bar{w} \leq w^{*T} Q w^*$ and $-\rho^T \bar{w} < -\rho^T w^* < 0$ then we derive that $\frac{\bar{w}^T Q \bar{w}}{\rho^T \bar{w}} < \frac{w^{*T} Q w^*}{\rho^T w^*}$. It contradicts with the Pareto optimality of w^*. The desired conclusion is followed. □

The result of Proposition 3.1 shows that if we find the Pareto set successfully, then it covers the solution given by Wang et al.[30] and therefore can improve the quality of feature selection. To solve Problem (SFS-MO), one can apply MPG-Explicit (Algorithm 1) by rewriting (SFS-MO) in the equivalent form

$$\min_{w \in \mathbb{R}^p} \quad \begin{cases} f_1(w) = g_1(w) + i_C(w) \\ f_2(w) = g_2(w) + i_C(w) \end{cases}, \tag{3.1}$$

where the indicator function of $C = \{w \in \mathbb{R}^p \mid \sum_{i=1}^{p} w_i = 1, w \geq 0\}$ defined by

$$i_C(w) = \begin{cases} 0, & \text{if } w \in C \\ +\infty, & \text{otherwise} \end{cases}.$$

The subproblem corresponding to Problem (2.13) now becomes

$$\min_{d \in \mathbb{R}^p} \quad \max_{i=1,2} \left\{ \nabla g_i(x^k)^T d + i_C(x^k + d) - i_C(x^k) \right\} + \frac{1}{2\alpha} \|d\|^2. \tag{3.2}$$

By introducing an auxiliary variable $\tau = \max\limits_{i=1,2} \nabla g_i(x^k)^T d$ and setting $d = u - x^k$, problem (3.2) can be rewritten as a quadratic programming as below:

$$\begin{aligned} \min_{u,\tau} \quad & \tau + \frac{1}{2\alpha}\|u - x^k\|^2 \\ \text{subject to} \quad & \nabla g_i(x^k)^T(u - x^k) \leq \tau, \quad i = 1, 2, \\ & \sum_{i=1}^{p} u_i = 1, u \geq 0. \end{aligned} \tag{3.3}$$

3.2. **Our proposed algorithm (Scaled-MPG-E) for solving Problem** (SFS-MO)**.** It is observed that Problem (SFS-MO) may have an imbalance property due to the different structure of the two objectives. This imbalance induces several disadvantages during the optimization process. Specifically, when the "redundancy minimization" objective spans an excessively large range of values, it dominates and prevails over the "relevancy maximization" objective. Consequently, the algorithm fails to distribute attention evenly across objectives, leading to suboptimal performance [14]. Furthermore, this scaling imbalance causes the obtained solutions to be unevenly distributed and show poor convergence [14]. Rather than achieving comprehensive coverage of the Pareto optimal set, the solutions tend to cluster and concentrate within a single region of the objective space [14].

In this section, we combine MPG-Explicit with the scaling technique given in [14] to propose the Scaled version of MPG-Explicit. Specifically, the objective functions are normalized according to the following formula:

$$\tilde{g}_i(x) = \frac{g_i(x) - z_i^{ideal}}{z_i^{nadir} - z_i^{ideal}}, \quad i = 1, 2, \tag{3.4}$$

where z^{ideal} and z^{nadir} are the ideal point and nadir point of Problem (SFS-MO), respectively. From [11], these points can be determined explicitly as presented in Procedure 2.

Procedure 2 Determination of Ideal and Nadir Points [11]

Step 1: Compute:

$$x_1^* = \arg\min_{x \in C} g_1(x) \quad \text{and} \quad x_2^* = \arg\min_{x \in C} g_2(x) \tag{3.5}$$

Step 2: Set:

$$z^{ideal} = \big(g_1(x_1^*), g_2(x_2^*)\big) \tag{3.6}$$

and

$$z^{nadir} = \big(g_1(x_2^*), g_2(x_1^*)\big) \tag{3.7}$$

Now, we obtain the scaled version of problem (SFS-MO) as follows

$$\begin{aligned} \min_{w} \quad & (\tilde{g}_1(w), \tilde{g}_2(w)) \\ \text{subject to} \quad & \sum_{i=1}^{p} w_i = 1, \\ & w \geq 0, \end{aligned} \tag{Scaled-SFS-MO}$$

which is equivalent to

$$\min_{w \in \mathbb{R}^p} \begin{cases} \tilde{f}_1(w) = \tilde{g}_1(w) + i_C(w) \\ \tilde{f}_2(w) = \tilde{g}_2(w) + i_C(w) \end{cases}. \tag{3.8}$$

The following proposition demonstrates the equivalence between Problem (Scaled-SFS-MO) and Problem (SFS-MO).

Proposition 3.2. *A point w^* is a Pareto optimal solution to problem* (Scaled-SFS-MO) *if and only if it is a Pareto optimal solution to problem* (SFS-MO).

Proof. Suppose that w^* is a Pareto optimal solution to problem (SFS-MO) then there does not exist $\bar{w} \neq w^*$ such that $g_i(\bar{w}) \leq g_i(w^*)$ for all $i \in \{1,2\}$, and $g_j(\bar{w}) < g_j(w^*)$ for at least one $j \in \{1,2\}$. Since the denominator $(z_i^{nadir} - z_i^{ideal})$ is positive for each i, hence there does not exist $\bar{w} \neq w^*$ such that

$$\tilde{g}_i(\bar{w}) = \frac{g_i(\bar{w}) - z_i^{ideal}}{z_i^{nadir} - z_i^{ideal}} \leq \frac{g_i(w^*) - z_i^{ideal}}{z_i^{nadir} - z_i^{ideal}} = \tilde{g}_i(w^*). \tag{3.9}$$

and $\tilde{g}_j(\bar{w}) < \tilde{g}_j(w^*)$ for at least one $j \in \{1,2\}$. This implies that w^* is also a Pareto optimal solution to problem (Scaled-SFS-MO). The "only if" statement is proved similarly.

□

Algorithm 2 Scaled-MPG-E

Step 1: Compute z^{ideal} and z^{nadir} by applying Procedure 2:

$$z^{ideal} = \big(g_1(w_1^*), g_2(w_2^*)\big) \quad \text{and} \quad z^{nadir} = \big(g_1(w_2^*), g_2(w_1^*)\big). \tag{3.10}$$

Step 2: Applying Algorithm 1 to Problem (Scaled-SFS-MO).

3.3. **An application to supervised feature selection problems.** The overall architecture of our proposed methodology is illustrated in Figure 1. Specifically, Figure 1a) presents an overview of the steps required to select the optimal feature set, while Figure 1b) provides a detailed description of the approach used to solve the multi-objective optimization problem (SFS-MO). Regarding the overall framework 1a): The matrix Q and vector ρ, computed from the features and data labels, serve as the input for solving the optimization problem (SFS-MO). To obtain the optimal feature subset, the optimal solutions—namely the score vectors w^* resulting from (SFS-MO) are processed through the *k-Winners-Take-All* (kWTA) neural network proposed by Wang et al.[27]. This transformation yields the binary vector z^*, whose components (0 or 1) represent the inclusion or exclusion of a specific feature in the final optimal subset.

4. Experimental results

Experiments were performed on a system featuring an Intel Core i5-1135G7 processor, 16.0 GB RAM, and integrated Intel(R) UHD Graphics. All source code was implemented in Python 3.14 and can be accessed at: https://github.com/hoaiphamthi/Scaled-MPG-E-for-SFS-MO.

4.1. **Performance analysis on clinical benchmark datasets.** The performance of the proposed methods (including NSGA-II, MPG-Explicit, and Scaled-MPG-E within a multi-objective framework) is evaluated through a comparative analysis against popular feature selection techniques, including Information Gain (IG), mRMR, CIFE, and the PNN technique proposed by Wang et al. [30].

In the experiments, we utilize 9 standard medical datasets to verify the reliability of the proposed algorithms across diverse diagnostic and biological classification tasks, including:

- *Neurological disorders:* Parkinsons and Darwin.
- *Oncological/Cardiovascular studies:* WDBC, Heart, and Arrhythmia.
- *Biochemical and Socio-medical analyses:* Gallstone, Blood, Musk-v1 and Toxic.

As summarized in Table 1, this experimental suite exhibits significant diversity in data scale, ranging from low-dimensional clinical records (12 features) to high-dimensional spaces (up to 1203 features). Furthermore, to estimate probability density functions within these benchmarks, we utilize histogram estimators with a fixed bin width. The number of bins is systematically determined by the formula $\frac{\log_2(n)}{2}$, where n represents the sample size of each dataset.

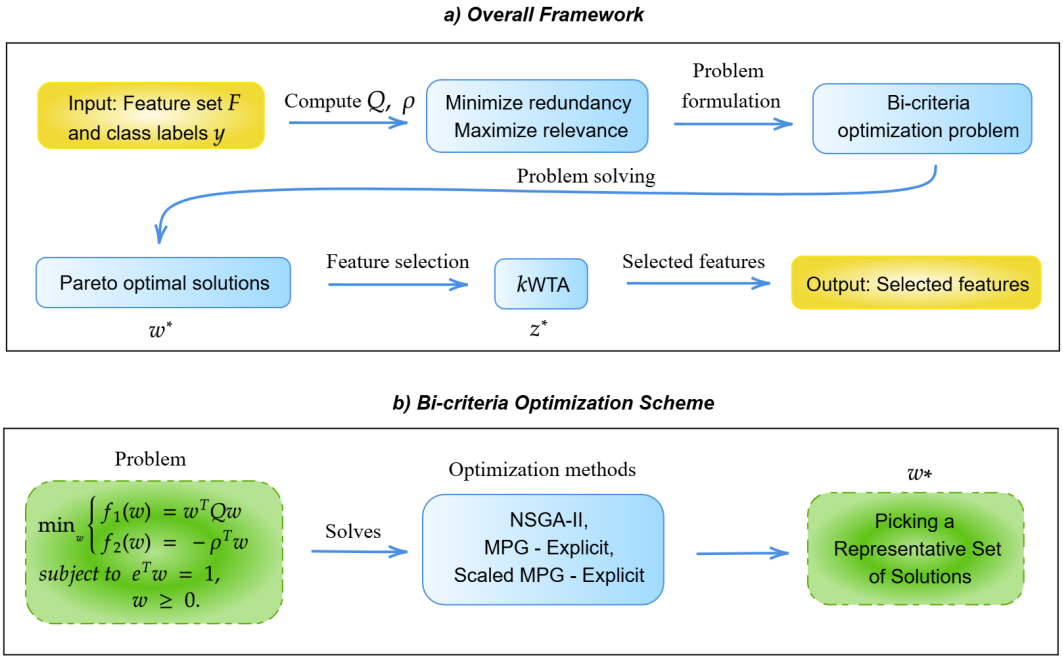

FIGURE 1. Proposed feature selection framework

TABLE 1. Nine benchmark datasets used in the experiments

Dataset	Features	Samples	Classes
Heart	12	270	2
Parkinsons	22	195	2
WDBC	30	569	2
Blood	32	5880	19
Gallstone	38	319	2
Musk-version 1	167	476	2
Arrhythmia	278	452	2
Darwin	450	174	2
Toxicity	1203	171	2

In terms of implementation, traditional feature selection techniques such as IG, mRMR, and CIFE are established using the `scikit-learn`, `mrmr-selection`, and `skfeature-chappers` libraries. The NSGA-II algorithm is configured using the `pymoo` library, while PNN and MPG-Explicit are constructed following the design methodologies of Wang et al. [30] and Bello-Cruz et al. [4], respectively, in which the quadratic programming subproblem (3.3) of the MPG-Explicit method is solved using the `cvxopt` library. For Scaled-MPG-E, `cvxopt` is also utilized to solve (3.5) to calculate the ideal and nadir points for the normalization process. These multi-objective methods are set up to search for a set of 250 Pareto-optimal points with a limit of 1000 iterations, while NSGA-II specifically is extended to 3000 iterations to ensure convergence.

After determining the score vectors, the process of extracting the subset of the k best features is conducted with sizes varying from 5 to 30 (in increments of 5) to analyze the trajectory of classification performance. Investigating this range of values not only helps evaluate the model's ability to maintain accuracy but also confirms the stability of the feature ranking, thereby ensuring that users can flexibly select the number of features based on practical needs while still obtaining significant attributes and consistent classification results.

The selected feature subsets are trained using two classification models: SVM and RF. Classification accuracy is adopted as the common metric for evaluation. The average classification results for each dataset across these two models are presented in Table 2 and Table 3, where the highest accuracy value for each dataset is highlighted in bold red, and the second-highest value is underlined in blue. The average classification accuracy of these two models across different numbers of selected features is illustrated via the line plot in Figure 2.

TABLE 2. Classification accuracy of SVM model on nine benchmark datasets

Dataset	Method							
	NSGA-II	MPG-E	S-MPG-E	PNN	IG	CIFE	mRMR	Baseline
Arrhythmia	0.7839	0.7363	0.7399	0.7381	0.7436	0.6722	0.7692	**0.7912**
Blood Cell	**0.9749**	0.9688	0.9685	0.9580	0.9473	0.8899	0.9483	0.9745
Darwin	**0.9143**	0.8238	0.8429	0.7524	0.8000	0.6667	0.8286	0.8286
Gallstone	**0.8307**	0.7578	0.7839	0.7370	0.7474	0.6380	0.7630	0.7344
Heart	**0.9074**	0.8519	0.8889	0.8519	0.8426	0.8056	0.8426	0.8519
Musk	0.8299	0.7257	0.7483	0.7361	0.7326	0.7448	0.7674	**0.8333**
Parkinsons	**0.9231**	0.8910	0.9167	0.8974	0.9038	0.9167	0.9103	0.8974
Toxicity	**0.7143**	0.6762	0.6762	0.6667	0.6762	0.6810	0.6476	0.6571
WDBC	**0.9825**	0.9722	0.9766	0.9708	0.9693	0.9459	0.9664	**0.9825**
Average	**0.8734**	0.8226	0.8380	0.8120	0.8181	0.7734	0.8270	0.8390

The experimental results synthesized from Table 2 and Table 3 show that the multi-objective approach outperforms the three traditional feature selection techniques (IG, mRMR, and CIFE) and PNN. In both models, NSGA-II consistently proves to be the best method, yielding superior results compared to the baseline model while significantly reducing the number of features. Notably, with the Toxicity dataset containing over 1,200 features, NSGA-II improves classification results by 5.72% – 8.58% even when the number of selected features does not exceed 30.

Following NSGA-II, Scaled-MPG-E delivers better results than all remaining feature selection methods. Furthermore, the line plots in Figure 2 further demonstrate the superior performance of NSGA-II, while MPG-Explicit and Scaled-MPG-E show highly competitive performance against popular feature selection methods.

To compare the solution quality of NSGA-II, MPG-Explicit, and Scaled-MPG-E, the obtained solutions are evaluated based on execution time, the number of non-dominated points, Hypervolume (HV) [34], and Inverted Generational Distance (IGD) [8]. It should be noted that the HV and IGD metrics in this study are calculated based on the local non-dominated set. The evaluation results are summarized in Table 4, where the best indicator for each dataset is highlighted in bold red. The shapes of the Pareto fronts for these three methods are also modeled in Figure 3 to clearly observe the distribution of the solutions.

TABLE 3. Classification accuracy of RF model on nine benchmark datasets

Dataset	Method							
	NSGA-II	MPG-E	S-MPG-E	PNN	IG	CIFE	mRMR	Baseline
Arrhythmia	0.7729	0.7179	0.7363	0.7253	0.7491	0.6813	0.7363	**0.8132**
Blood Cell	0.9722	0.9673	0.9671	0.9517	0.9435	0.9019	0.9418	**0.9762**
Darwin	**0.9048**	0.8381	0.8429	0.7619	0.8333	0.7095	0.8667	0.8857
Gallstone	**0.8776**	0.8229	0.8385	0.8099	0.8177	0.6354	0.8281	0.7969
Heart	**0.8796**	0.8148	0.8704	0.8148	0.7778	0.7870	0.8241	0.8333
Musk	**0.8333**	0.7674	0.7708	0.7639	0.7656	0.8281	0.7882	0.8229
Parkinsons	**0.9487**	0.9359	0.9423	0.9231	0.9359	0.9423	0.9423	**0.9487**
Toxicity	**0.7429**	0.6857	0.7095	0.6333	0.6429	0.5714	0.6286	0.6571
WDBC	**0.9737**	0.9620	0.9635	0.9620	0.9591	0.9488	0.9605	0.9649
Average	**0.8784**	0.8347	0.8490	0.8162	0.8250	0.7784	0.8352	0.8554

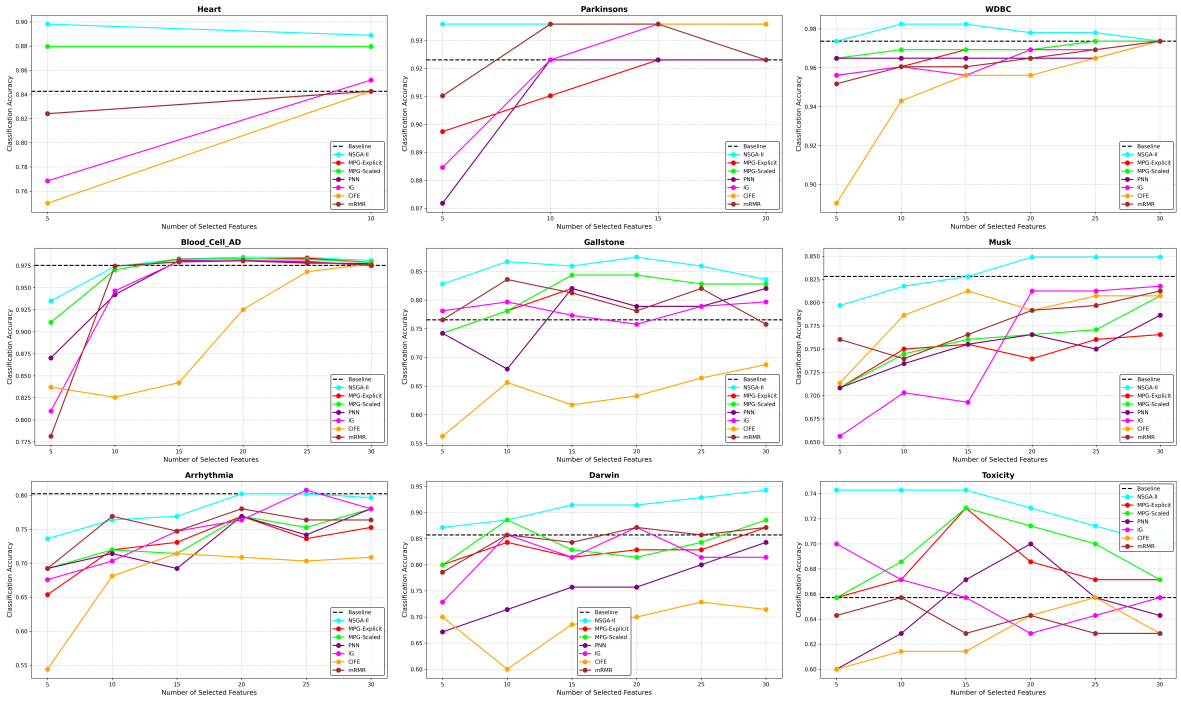

FIGURE 2. Average classification accuracy of SVM and RF models across varying numbers of selected features.

According to Table 4, the Hypervolume and IGD indicators for NSGA-II are superior to those of both MPG-Explicit and Scaled-MPG-E. Furthermore, observations from Figure 3 reveal that the Pareto front of NSGA-II is significantly broader and more evenly distributed than those of MPG-Explicit and Scaled-MPG-E. This demonstrates the superior capability of NSGA-II in identifying solutions that cover the objective space. Due to its extensive search range, NSGA-II is able to identify w^* points for higher-quality feature subsets, resulting in better classification accuracy. However, NSGA-II yields a relatively

TABLE 4. Performance comparison of multi-objective methods based on quality metrics and execution time.

Dataset	Method	Time (s)	G_Nondom	HV	IGD
Arrhythmia	MPG-E	553.58	248	0.0660	0.0410
	S-MPG-E	**117.03**	**250**	0.0687	0.0270
	NSGA-II	185.26	152	**0.0688**	**0.0020**
Blood Cell	MPG-E	**32.19**	**250**	1.2348	0.0090
	S-MPG-E	36.19	**250**	1.2357	0.0088
	NSGA-II	47.75	99	**1.2395**	**0.0026**
Darwin	MPG-E	867.67	**249**	0.1305	0.0218
	S-MPG-E	**315.88**	248	0.1331	0.0171
	NSGA-II	352.13	144	**0.1340**	**0.0030**
Gallstone	MPG-E	236.27	248	0.0622	0.0168
	S-MPG-E	**36.44**	**250**	0.0640	0.0109
	NSGA-II	49.57	144	**0.0662**	**0.0012**
Heart	MPG-E	63.67	**250**	0.1121	0.0176
	S-MPG-E	**19.01**	249	0.1148	0.0099
	NSGA-II	41.33	146	**0.1153**	**0.0009**
Musk	MPG-E	296.08	249	0.0897	0.0211
	S-MPG-E	**68.36**	**250**	0.0937	0.0114
	NSGA-II	108.63	130	**0.0951**	**0.0019**
Parkinsons	MPG-E	67.24	**250**	0.1493	0.0163
	S-MPG-E	**27.18**	**250**	0.1525	0.0137
	NSGA-II	44.07	132	**0.1544**	**0.0010**
Toxicity	MPG-E	8646.90	**246**	0.0161	0.0163
	S-MPG-E	2923.43	243	0.1751	0.0053
	NSGA-II	**2032.13**	118	**0.0177**	**0.0025**
WDBC	MPG-E	54.89	**250**	0.3942	0.0243
	S-MPG-E	**30.76**	**250**	0.4005	0.0157
	NSGA-II	46.44	142	**0.4023**	**0.0014**

low number of non-dominated solutions compared to the other two methods, suggesting that the algorithm has not yet achieved deep convergence.

Scaled-MPG-E leads in terms of execution time; furthermore, the metrics for non-dominated points, HV, and IGD all indicate that Scaled-MPG-E offers a significant performance improvement over the original MPG-Explicit. Although it follows NSGA-II in classification accuracy across the experimental

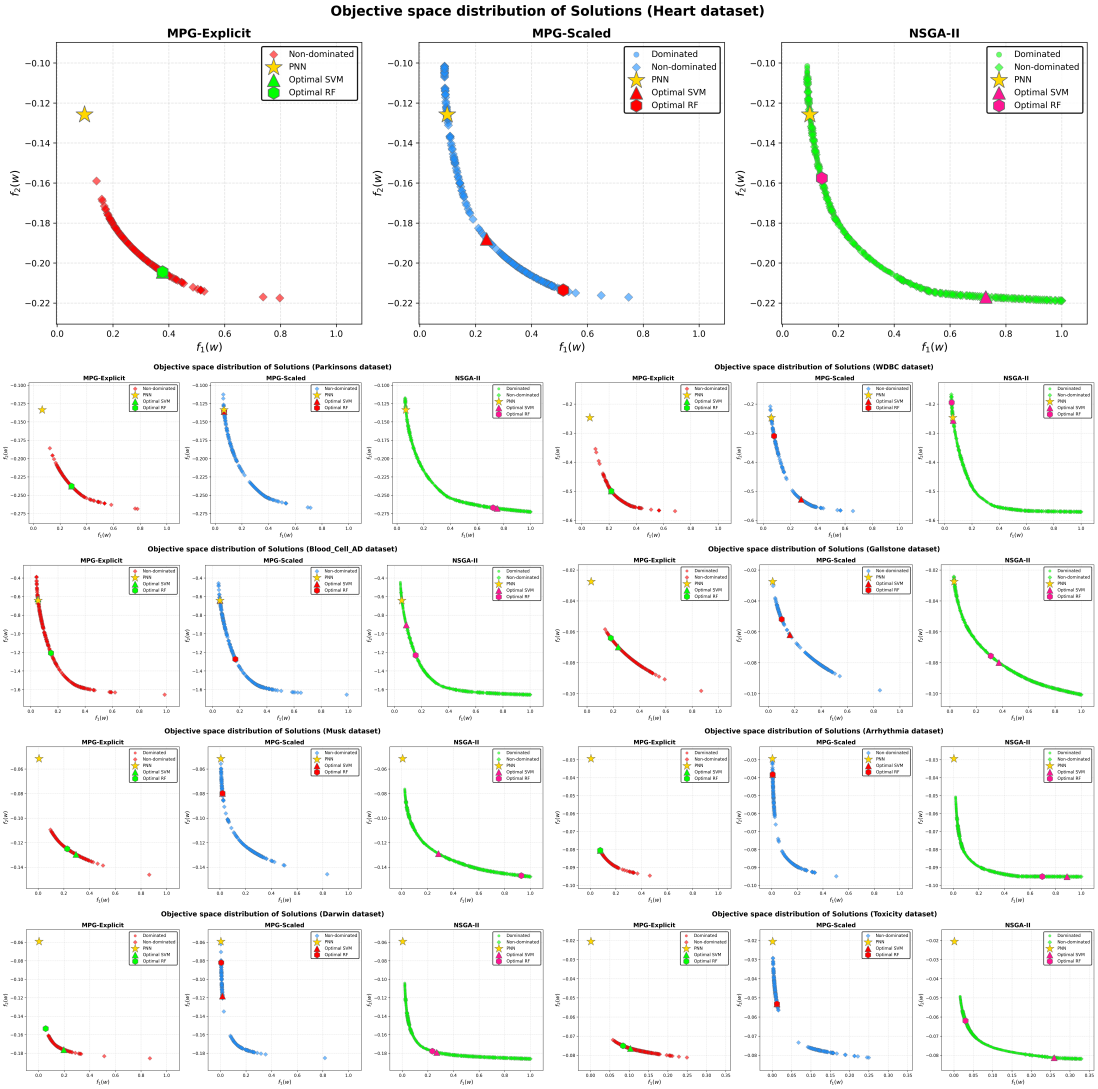

FIGURE 3. Visualization of Pareto fronts for NSGA-II, MPG-Explicit, and Scaled-MPG-E across datasets.

datasets, Scaled-MPG-E still demonstrates more outstanding performance than traditional feature selection methods. Combined with its strength in processing speed, Scaled-MPG-E is well-suited for medical decision support systems that require a balance between performance and practical processing time.

4.2. **Application in skin cancer classification and clinical decision support.** In this section, we evaluate the effectiveness of the proposed method in the diagnostic classification of skin cancer: benign versus malignant. Experiments were conducted on the HAM10000 dataset (https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000, Figure 4), which comprises over 10,015 multisource dermatoscopic images of common pigmented lesions along with their corresponding masks (https://www.kaggle.com/datasets/tschandl/ham10000-lesion-segmentations, Figure 5). The labels within this dataset were validated by medical experts through biopsy or direct clinical diagnosis, ensuring the ground-truth accuracy for the model training and validation processes.

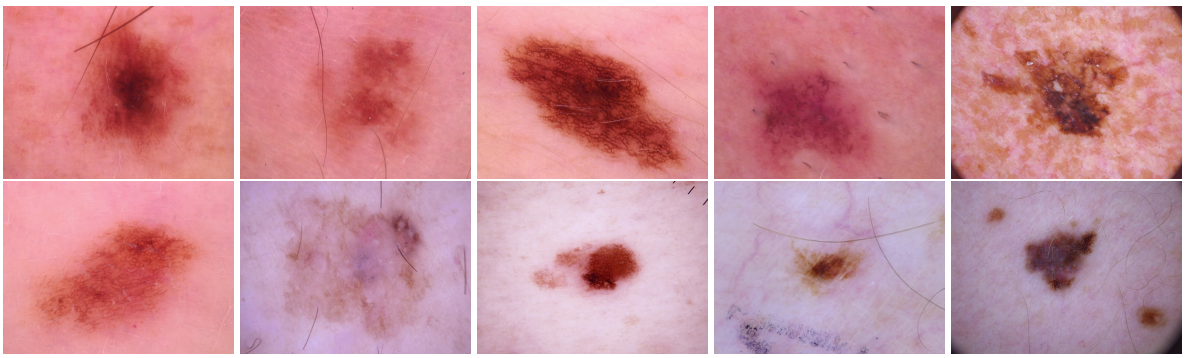

FIGURE 4. Examples of skin lesion images from the HAM10000 dataset

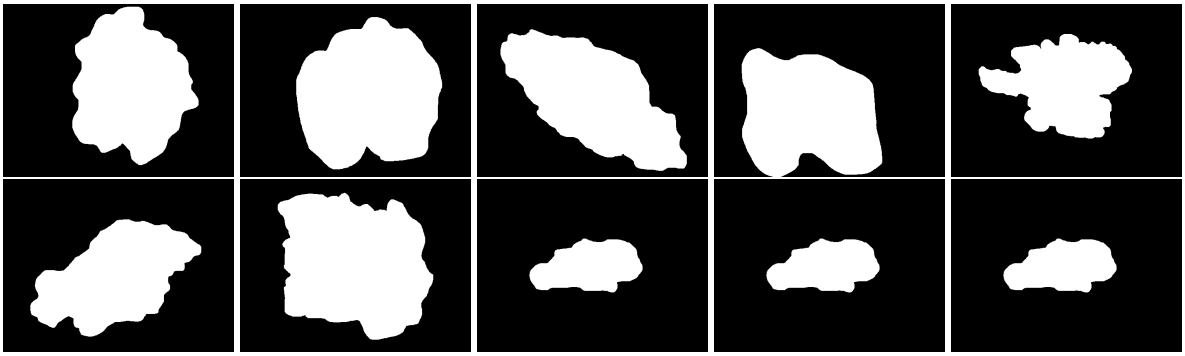

FIGURE 5. Examples of corresponding ground-truth masks from the HAM10000 dataset

To perform this classification task, we manually extract features with practical medical significance, such as color, shape, and pigment asymmetry, along with several features from the accompanying metadata. A total of 366 input features are utilized for this problem, as listed in Table 5.

TABLE 5. Detailed description of the 366 extracted handcrafted features.

Feature group	Quantity	Description
Morphological	48	Shape and boundary descriptors including area, perimeter, asymmetry, convex-hull properties, Hu moments, and Zernike moments
Colorimetric	128	Statistical, entropy, variance, range, and histogram features across HSV & LAB color spaces
Textural	166	Surface pattern descriptors including GLCM, Haralick, Tamura, edge, LBP, Gabor, HOG, and wavelet features
Frequency	20	Fourier descriptors from contour analysis
Clinical Metadata	4	Patient information including Age, Sex, Anatomical site, and Diagnosis type

The dataset exhibits a slight imbalance with a benign-to-cancer ratio of approximately 4:1. During implementation, we addressed this issue using the SMOTE-Tomek technique from Python's `imbalanced-learn` library [3]. Furthermore, we will evaluate the proposed multi-objective methods alongside the PNN technique proposed by Wang et al. [30] and compare their performance when coupled with the SVM model.

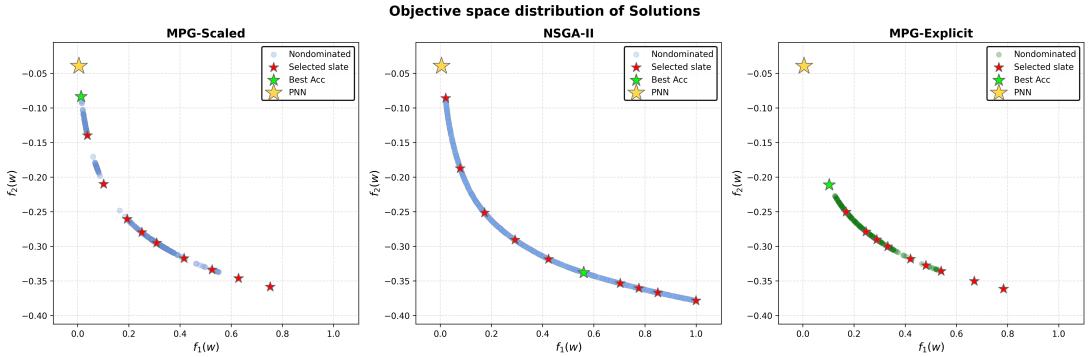

FIGURE 6. Distribution of the 10 representative Pareto points selected by the coverage method

A fundamental question arises regarding how to reduce the computational cost as the scale of the feature selection problem increases significantly. To address this challenge, we employ the coverage method proposed by Sayin et al. [22] to extract a representative subset from the initial Pareto optimal set. Instead of evaluating the entire large set of Pareto points, we only need to assess this representative subset, thereby saving resources and substantially reducing computation time.

Specifically, in this experiment, we identify a set of 500 Pareto points and subsequently select a subset containing 10 representative points for the Pareto front. As observed in Figure 6, the selected Pareto points demonstrate a uniform coverage across the entire frontier. These points do not suffer from clustering in high-density regions; instead, they ensure a broad distribution from the beginning to the end of the Pareto front, even in regions where solutions are less concentrated.

The experimental results presented in Table 6 confirm the efficacy of the multi-objective optimization methods. Specifically, the S-MPG-E algorithm not only achieves higher classification accuracy than the baseline model but also significantly optimizes the input space by reducing the number of features by 61.75%. The effective removal of noise components allows the SVM model to focus on extracting the most decisive features, thereby enhancing overall performance. Simultaneously, this improvement is reflected through the AUC metrics and ROC curves in Figure 7. Furthermore, other multi-objective methods also record results that closely follow the performance of MPG-Explicit. Overall, the multi-objective methods demonstrate greater effectiveness compared to the PNN method.

TABLE 6. Classification results and feature selection efficiency of different methods on the skin cancer dataset

	Baseline	S-MPG-E	NSGA-II	MPG-E	PNN
No. of features	366	**140**	160	145	145
Accuracy	0.8822	**0.8952**	0.8812	0.8947	0.8937
AUC	0.9356	**0.9463**	0.9365	0.9453	0.9445

The quality metrics for the solution sets of the optimization algorithms—including the number of non-dominated points, Hypervolume (HV) [34], Inverted Generational Distance (IGD) [8], and computational time are summarized in Table 7. It should be noted that the HV and IGD metrics in this study are calculated based on the local non-dominated set.

The results in Table 7, together with the Pareto front distribution in Figure 6, reaffirm that the Scaled-MPG-E algorithm not only significantly optimizes execution time but also substantially expands the coverage of the obtained Pareto front. However, an important observation is that although Proposition

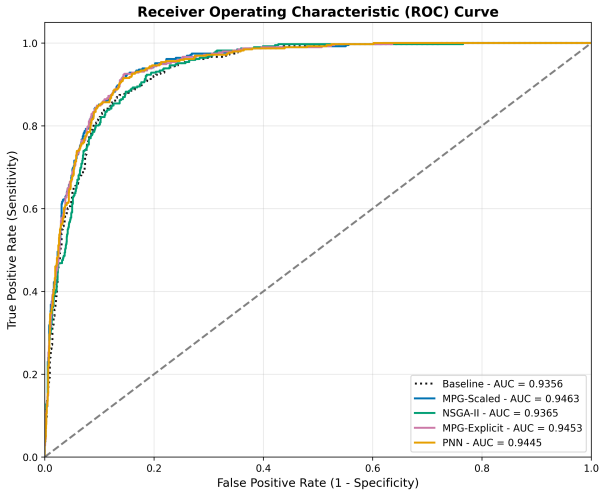

FIGURE 7. ROC curves of various methods on the skin cancer dataset

3.1 proves that the point found by the PNN method belongs to the Pareto front, the experimental results in Figure 6 show that none of the three obtained Pareto fronts reach the position of this PNN point. This is a notable limitation, indicating a gap that needs to be addressed in current multi-objective optimization algorithms.

TABLE 7. Comparative performance analysis between the full feature set (Baseline) and the MPG-Explicit selected subset

	Time (s)	G_Nondom	HV	IGD
MPG-E	586.2	**500**	0.2496	0.0174
S-MPG-E	**275.8**	499	0.2552	0.0117
NSGA-II	692.7	269	**0.2621**	**0.0009**

5. CONCLUSIONS

In this study, we have proposed an effective multi-objective approach for the supervised feature selection problem. Its efficient performance is validated through direct comparisons with existing methods on various data sets in medicine. Furthermore, our new algorithm (Scaled-MPG-E), designed based on the integration of a scaling technique into the exact algorithm (MPG-Explicit), significantly accelerates processing time and expands the Pareto front compared to the original approach. Future research could study how to improve the diversity of Pareto fronts and investigate the impact of objective function disparity on the performance of exact algorithms.

STATEMENTS AND DECLARATIONS

The authors declare that they have no conflict of interest. The the manuscript has associated data in this link https://github.com/hoaiphamthi/Scaled-MPG-E-for-SFS-MO.

REFERENCES

[1] J. C. Ang, A. Mirzal, H. Haron, and H. N. Hamed. Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 13(5):971–989, 2016. https://doi.org/10.1109/TCBB.2015.2478454.

[2] B. Azhagusundari and A. S. Thanamani. Feature selection based on information gain. *Int. J. Innov. Technol. Explor. Eng.*, 2(2):18–21, 2013.

[3] G. E. Batista, R. C. Prati, and M. C. Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29, 2004. https://doi.org/10.1145/1007730.1007735.

[4] Y. Bello-Cruz, J. Melo, L. Prudente, and R. Serra. A proximal gradient method with an explicit line search for multiobjective optimization. *Computational Optimization and Applications*, 92:437–469, 2025. https://doi.org/10.1007/s10589-025-00711-x.

[5] A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1):245–271, 1997. https://doi.org/10.1016/S0004-3702(97)00063-5.

[6] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *J. Mach. Learn. Res.*, 13(Jan):27–66, 2012. https://jmlr.org/papers/v13/brown12a.html.

[7] T. M. Cover and J. A. Thomas. *Elements of Information Theory*, volume 2nd edition. John Wiley & Sons, Hoboken, New Jersey, 2nd edition, 2006. https://doi.org/10.1002/047174882X.

[8] C. Dai, Y. Wang, and M. Ye. A new multi-objective particle swarm optimization algorithm based on decomposition. *Applied Soft Computing*, 30:384–397, 2015. https://doi.org/10.1016/j.asoc.2015.01.037.

[9] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002. https://doi.org/10.1109/4235.996017.

[10] R. O. Duda, P. E. Hart, et al. *Pattern classification*. John Wiley & Sons, 2006.

[11] M. Ehrgott and D. Tenfelde-Podehl. Computation of ideal and nadir values and implications for their use in mcdm methods. *European Journal of Operational Research*, 151(1):119–139, 2003. https://doi.org/10.1016/S0377-2217(02)00595-7.

[12] B. Ghojogh, M. N. Samad, S. A. Mashhadi, T. Kapoor, W. Ali, F. Karray, and M. Crowley. Feature selection and feature extraction in pattern analysis: A literature review, 2019. https://arxiv.org/abs/1905.02845.

[13] J. Gui, Z. Sun, S. Ji, D. Tao, and T. Tan. Feature selection based on structured sparsity: A comprehensive study. *IEEE Transactions on Neural Networks and Learning Systems*, 28(7):1490–1507, 2016. https://doi.org/10.1109/TNNLS.2016.2555866.

[14] L. He, H. Ishibuchi, A. Trivedi, H. Wang, Y. Nan, and D. Srinivasan. A survey of normalization methods in multiobjective evolutionary algorithms. *IEEE Transactions on Evolutionary Computation*, 2023. https://ieeexplore.ieee.org/document/9419072.

[15] P. T. Hoai, N. D. Hoang, and F. Lara. A modified projected gradient algorithm for solving quasiconvex programming with applications. 2026. https://optimization-online.org/?p=33877.

[16] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1):273–324, 1997. https://doi.org/10.1016/S0004-3702(97)00043-X.

[17] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6):1–45, 2017. https://doi.org/10.1145/3136625.

[18] D. Lin and X. Tang. Conditional infomax learning: An integrated framework for feature extraction and fusion. In A. Leonardis, H. Bischof, and A. Pinz, editors, *Computer Vision – ECCV 2006*, pages 68–82, Berlin, Heidelberg, 2006. https://doi.org/10.1007/11744023_6. Springer Berlin Heidelberg.

[19] F. Nie, S. Yang, R. Zhang, and X. Li. A general framework for auto-weighted feature selection via global redundancy minimization. *IEEE Transactions on Image Processing*, 28(5):2428–2438, 2018. https://doi.org/10.1109/TIP.2018.2886761.

[20] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005. https://doi.org/10.1109/TPAMI.2005.159.

[21] I. Rodriguez-Lujan, R. Ramon-Cano, Y. Iturria-Medina, D. Martinez-Rego, E. Jimenez-Hernandez, M. Heredia-Conde, and A. Tonda. Quadratic programming feature selection. *Journal of Machine Learning Research*, 11:1491–1516, 2010. https://jmlr.org/papers/v11/rodriguez-lujan10a.html.

[22] S. Sayın. Measuring the quality of discrete representations of efficient sets in multiple objective mathematical programming. *Mathematical Programming*, 87(3):543–560, 2000. https://link.springer.com/article/10.1007/s101070050011.

[23] N. Srinivas and K. Deb. Multiobjective optimization using nondominated sorting in genetic algorithms. *Evolutionary computation*, 2(3):221–248, 1994. https://doi.org/10.1162/evco.1994.2.3.221.

[24] M. Tan, L. Wang, and I. W. Tsang. Feature-selected tree-based classification. *IEEE Transactions on Cybernetics*, 43(6):1990–2004, 2013. https://doi.org/10.1109/TSMCB.2012.2237394.

[25] H. Tanabe, E. H. Fukuda, and N. Yamashita. Proximal gradient methods for multiobjective optimization and their applications. *Computational Optimization and Applications*, 72(2):339–361, 2019. https://doi.org/10.1007/s10589-018-0043-x.

[26] S. Verma, M. Pant, and V. Snasel. A comprehensive review on NSGA-II for multi-objective combinatorial optimization problems. *IEEE Access*, 9:82675–82791, 2021. https://doi.org/10.1109/ACCESS.2021.3086364.

[27] J. Wang. Analysis and design of a k -winners-take-all model with a single state variable and the heaviside step activation function. *IEEE Transactions on Neural Networks*, 21(9):1496–1506, 2010. https://doi.org/10.1109/TNN.2010.2052631.

[28] X. Wang, Y. Liu, F. Nie, and H. Huang. Discriminative unsupervised dimensionality reduction. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI '15)*, pages 3925–3931. AAAI Press, 2015. https://www.ijcai.org/Proceedings/15/Papers/552.pdf.

[29] Y. Wang, X. Li, and R. Ruiz. Weighted general group lasso for gene selection in cancer classification. *IEEE Transactions on Cybernetics*, 49(8):2860–2873, 2019. https://doi.org/10.1109/TCYB.2018.2829811.

[30] Y. Wang, X. Li, and J. Wang. A neurodynamic optimization approach to supervised feature selection via fractional programming. *Neural Networks*, 136:194–206, 2021. https://doi.org/10.1016/j.neunet.2021.01.004.

[31] H. Yang and J. Moody. Feature selection based on joint mutual information. In *Proceedings of the International ICSC Symposium on Advances in Intelligent Data Analysis*, pages 22–25, 1999.

[32] R. W. Yeung. A new outlook on shannon's information measures. *IEEE Trans. Inf. Theory*, 37(3):466–474, 1991. https://doi.org/10.1109/18.79902.

[33] H. Zhang, J. Wang, Z. Sun, J. M. Zurada, and N. R. Pal. Feature selection for neural networks using group lasso regularization. *IEEE Transactions on Knowledge and Data Engineering*, 32(4):659–673, 2020. https://doi.org/10.1109/TKDE.2019.2893266.

[34] E. Zitzler, L. Thiele, M. Laumanns, C. M. Fonseca, and V. G. Da Fonseca. Performance assessment of multiobjective optimizers: An analysis and review. *IEEE Transactions on Evolutionary Computation*, 7(2):117–132, 2003. https://doi.org/10.1109/TEVC.2003.810758.