# A hybrid feature selection algorithm and its application in bioinformatics

Yangyang Wang [1], Xiaoguang Gao [1], Xinxin Ru [1], Pengzhan Sun [1], Jihan Wang [Corresp. 2]

[1] School of Electronics and Information, Northwestern Polytechnic University, Xi'an, Shaanxi, China

[2] Institute of Medical Research, Northwestern Polytechnic University, Xi'an, Shaanxi, China

Corresponding Author: Jihan Wang
Email address: jihanwang@nwpu.edu.cn

Feature selection is an independent technology for high-dimensional dataset that has been widely applied in a variety of fields. With the vast expansion of information, such as bioinformatics data, there has been an urgent need to investigate more effective and accurate methods involving feature selection in recent decades. Here, we proposed the hybrid MMPSO method, by combining the feature ranking method and the heuristic search method, to obtain an optimal subset that can be used for higher classification accuracy. In this study, ten datasets obtained from the UCI Machine Learning Repository were analyzed to demonstrate the superiority of our method. The MMPSO algorithm outperformed other algorithms in terms of classification accuracy while utilizing the same number of features. Then we applied the method to a biological dataset containing gene expression information about liver hepatocellular carcinoma (LIHC) samples obtained from The Cancer Genome Atlas (TCGA) and Genotype-Tissue Expression (GTEx). On the basis of the MMPSO algorithm, we identified a 18-gene signature that performed well in distinguishing normal samples from tumours. Nine of the 18 differentially expressed genes were significantly up-regulated in LIHC tumour samples, and the area under curves (AUC) of the combination seven genes (ADRA2B, ERAP2, NPC1L1, PLVAP, POMC, PYROXD2, TRIM29) in classifying tumours with normal samples was greater than 0.99. Six genes (ADRA2B, PYROXD2, CACHD1, FKBP1B, PRKD1 and RPL7AP6) were significantly correlated with survival time. The MMPSO algorithm can be used to effectively extract features from a high-dimensional dataset, which will provide new clues for identifying biomarkers or therapeutic targets from biological data and more perspectives in tumor research.

1 # A hybrid feature selection algorithm and its application in
2 # bioinformatics

4 Yangyang Wang[1], Xiaoguang Gao[1], Xinxin Ru[1], Pengzhan Sun[1], Jihan Wang[2]
5 [1] School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China
6 [2] Institute of Medical Research, Northwestern Polytechnical University, Xi'an, China

8 **Corresponding Author:**
9 Xiaoguang Gao[1], Jihan Wang[2]
10 [1] School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China.
11 Email: xggao@nwpu.edu.cn
12 [2] Institute of Medical Research, Northwestern Polytechnical University, Xi'an, China. Email:
13 jihanwang@nwpu.edu.cn

15 ## ABSTRACT

16 Feature selection is an independent technology for high-dimensional dataset that has been widely
17 applied in a variety of fields. With the vast expansion of information, such as bioinformatics
18 data, there has been an urgent need to investigate more effective and accurate methods involving
19 feature selection in recent decades. Here, we proposed the hybrid maximum information
20 coefficient, minimum-redundancy maximum-relevancy, and particle swarm optimization
21 (MMPSO) method, by combining the feature ranking method and the heuristic search method, to
22 obtain an optimal subset that can be used for higher classification accuracy. In this study, ten
23 datasets obtained from the UCI Machine Learning Repository were analyzed to demonstrate the
24 superiority of our method. The MMPSO algorithm outperformed other algorithms in terms of
25 classification accuracy while utilizing the same number of features. Then we applied the method
26 to a biological dataset containing gene expression information about liver hepatocellular
27 carcinoma (LIHC) samples obtained from The Cancer Genome Atlas (TCGA) and Genotype-
28 Tissue Expression (GTEx). On the basis of the MMPSO algorithm, we identified a 18-gene
29 signature that performed well in distinguishing normal samples from tumours. Nine of the 18
30 differentially expressed genes were significantly up-regulated in LIHC tumour samples ($P <$
31 $0.01$), and the area under curves (AUC) of the combination seven genes (ADRA2B, ERAP2,
32 NPC1L1, PLVAP, POMC, PYROXD2, TRIM29) in classifying tumours with normal samples
33 was greater than $0.99$. Six genes (ADRA2B, PYROXD2, CACHD1, FKBP1B, PRKD1 and
34 RPL7AP6) were significantly correlated with survival time ($P < 0.05$). Taken together, the
35 MMPSO algorithm can be used to effectively extract features from a high-dimensional dataset,
36 which will provide new clues for identifying biomarkers or therapeutic targets from biological
37 data and more perspectives in tumor research.
38 **Keywords:** Feature selection; MMPSO; Biomarkers; ROC; Survival analysis

39

## INTRODUCTION

40  **INTRODUCTION**
41  The dimensionality of data has increased greatly due to the rapid growth in big data(Li et al.
42  2020; Wainwright 2019). This condition has also accelerated the development of high
43  dimensional data processing technology(Li et al. 2016; Saeys et al. 2007). One of the main issues
44  in data mining, pattern recognition, and machine learning is feature selection for high
45  dimensional data(Chen et al. 2020; Larranaga et al. 2006). Feature selection is the process of
46  selecting the feature subset that best captures the characteristics of the original dataset and alters
47  the feature expression of the original dataset as little as possible. It can be utilized as an
48  important dimensionality reduction technique to minimize computing complexity, lower the
49  potential of overfitting as well as improve the prediction performance(Tao et al. 2015). Feature
50  selection seldom modifies the original feature space, and the resultant feature subset has clearer
51  physical implications that can be exploited for subsequent classification or inference(Villa et
52  al. 2021). The search for the optimal subset of features is typically computationally expensive
53  and has been demonstrated to be nondeterministic polynomial-hard (NP-hard)(Faris et al. 2018;
54  Wang et al. 2016). Traditionally, feature selection algorithms are classified into three categories:
55  filter, wrapper, and embedded methods and these methods can also be divided into two main
56  categories: feature ranking and feature subset selection(Van Hulse et al. 2011). In the past few
57  years, feature selection based on high-dimensional datasets has attracted more attentions.
58  Because of their simplicity and efficiency, ranking-based approaches such as ReliefF (Robnik-
59  Šikonja & Kononenko 2003), minimum-redundancy maximum-relevancy (mRMR)(Peng et al.
60  2005), Fisher (Gu et al. 2012), CFS (Hong et al. 2011), and others are widely utilized in a variety
61  of applications. Different from the feature ranking selection, which screens out the top $K$ highest-
62  scoring features, feature subset selection selects the subset of features that perform well together.
63  Some heuristic search strategies(Rasheed & Education 2021) have been proposed to obtain the
64  global optimal feature subset, such as the genetic algorithm (GA)(Holland 1975; Stefano et al.
65  2017), particle swarm optimization (PSO)(Chuang et al. 2008; Eberhart & Kennedy 2002;
66  Xiangyang et al. 2007), and ant colony optimization (ACC)(Li et al. 2008a). It is worth
67  mentioning that, some methods based on neural networks which supports higher-dimensional
68  inputs can also be used for feature selection(Liu et al. 2022). Feature selection has been widely
69  utilized in bioinformatics to remove irrelevant features in high-throughput data as an effective
70  method for preventing the "curse of dimensionality"(Li et al. 2008b). It is appropriate to filter
71  out biomarkers in the medical field, which can not only help explore disease pathophysiology at
72  the molecular level but also has advantages in accurate diagnosis. In general, the number of
73  features in a bioinformatics dataset tends to be very large. It is critical to identify highly
74  discriminating biomarkers to improve disease diagnosis and prediction accuracy(Ma et al. 2020).
75  Therefore, there is no doubt that obtaining relevant biomarkers from high-throughput data is of
76  great significance (Yuanyuan et al. 2021). Furthermore, we realized that there is considerable
77  space for improvement in the feature selection process by combining feature ranking with feature
78  subset searching. There are several methods for measuring the specific value of relevance,
79  including the Pearson correlation coefficient (Obayashi & Kinoshita 2009), mutual information

80   (MI), and maximum information coefficient (MIC)(Reshef et al. 2011), and MIC can substitute
81   MI to obtain better mutual information measurement results in some situations, particularly for
82   continuous data. Furthermore, feature ranking does not provide a "golden standard" for obtaining
83   the best feature subset but only the ranking result. Therefore, combining the two methods is a
84   more promising way(Shreem et al. 2013; Stefano et al. 2017).
85     In this study, we focus on developing a hybrid efficient approach for obtaining the optimum
86   features by combining the feature ranking method and the heuristic search method. Specifically,
87   ten datasets were employed to validate our hypothesis first. Furthermore, one dataset derived
88   from high-throughput sequencing was used to assess the effect of the approach at the genetic
89   level. The discussion and conclusion are presented in the last section.
90

## MATERIALS & METHODS
**The mRMR algorithm**

The mRMR(Peng et al. 2005) algorithm, which uses mutual information to assess the relevance
between features, has been used in bioinformatics(Ding & Peng 2005; Li et al. 2012; Mundra &
Rajapakse 2010). Mutual information is widely used to analyze the correlation between two
variables, and it can be expressed as Equation 1.

$$I(X,Y) = \sum_{x \in X}\sum_{y \in Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \#(1)$$

In the Equation 1, P represents the probability and $X$, $Y$ represent the feature vector or class
vector. The relevance $V$ and redundancy $W$ of the mRMR can be expressed using Equation 2 and
Equation 3.

$$V = \frac{1}{|S|}\sum_{x_i \in S} I(y;x_i) \#(2)$$

$$W = \frac{1}{|S|^2}\sum_{x_i,x_j \in S} I(x_i;x_j) \#(3)$$

In the Equation 2 and 3, $y$ is the target variable, $S$ is candidate feature set and $x_i$, $x_j$ are arbitrary
variables of $S$. To calculate the final score of relevance, the $MIQ$ can be used, as shown in
Equation 4.

$$MIQ: argmax\left(\frac{V}{W}\right) \#(4)$$

**Maximal Information Coefficient**

As a measure of dependence for two-variable relationships in a large dataset, MIC has been
widely used in various fields, including global health, gene expression, human gut microbiota
and identify novel relationships due to its ability to capture a wide range of functional and non-
functional associations.  The definition of MIC is:

$$MIC(x,y \mid D) = max_{i*j < B(n)}\left\{\frac{I^*(x,y,D,i,j)}{\log \min(i,j)}\right\} \#(5)$$

115    In the Equation 5, $x$ and $y$ are the *pairs(x, y)* of the dataset and *I\*(x, y, D, i, j)* denotes the
116    maximum mutual information of $D|_G$ with the *i-by-j* grid, where the default $B(n) = n^{0.6}$.
117
118    **Particle swarm optimization algorithm**
119    The particle swarm optimization (PSO) algorithm is a heuristic search algorithm that originated
120    from studies of bird predation behavior(Eberhart & Kennedy 2002). The first step of PSO is to
121    initialize a group of particles. It then iterates until it finds the best solution. The particles update
122    themselves in each iteration by tracking two extreme values. The first is the particle's
123    determination of the individual extreme value *Pbest*. The other is called *Gbest*, which is
124    determined by the entire particle swarm. After determining the *Gbest* and *Pbest* values, the
125    particle updates its speed and position based on Equation 6 and 7.
126
127    $$V_{k+1} = \omega V_k + c_1 r_1 (Pbest - X_k) + c_2 r_2 (Gbest - X_k) \#(6)$$
128    $$X_{k+1} = X_k + V_k \#(7)$$
129
130    In the above equations, $k$ represents the number of iterations; $V_k$ *and* $X_k$ represent the particle's
131    current velocity and position, respectively; $r_1$, $r_2$ are random values between [0, 1]; $c_1$, $c_2$ are the
132    learning factors; $\omega$ is the inertia weight, which is used to control the influence of the last
133    iteration's speed on the current speed. A smaller and larger $\omega$ can strengthen the PSO
134    algorithm's local or global search ability, respectively.
135
136    **The hybrid algorithm for feature selection**
137    In this study, we proposed the MMPSO hybrid method, as shown in **Algorithm 1**. First, the
138    dataset needed to be preprocessed. On the one hand, the aim of preprocessing is to remove some
139    features that contain a large quantity of noisy data, such as features that contain many zeros. On
140    the other hand, if the proportion of samples is clearly unbalanced, it is necessary to balance the
141    samples. Here, we employed random oversampling technology to address this issue. A random
142    over-sampler randomly copies and repeats the minority class samples, eventually resulting in the
143    minority and majority classes having the same number. The next step was to rank the features.
144    MIC was used to measure the correlation of two features, resulting in a more accurate ranking
145    result of features based on the mRMR framework(Cao et al. 2021). Considering the numerous
146    features and the complexity of MIC, we used the multithreading method in paper(Tang et al.
147    2014) to speed up the calculation. After performing the mRMR based on MIC method, we
148    obtained the ranking features and use the top $K$ features as the input of the next step to reduce the
149    computational load for the PSO. The $K$ features were used in the third step to initialize the
150    particle swam and calculate the fitness of each particle. For the wrapper feature selection
151    algorithm, only the classification accuracy is used as the fitness function to guide the feature
152    selection process, which will lead to a larger scale of the selected feature subset(Liu et al. 2011).
153    Therefore, some studies combine the classification accuracy and the number of selected feature
154    subsets to form a fitness function(Xue et al. 2012). Here, the fitness we defined is shown in
155    Equation 8. $V_{error}$ was the error, which is measured by a classification method of k-nearest
156    neighbor (KNN)(Chen et al. 2021a). $N_{selected}$ and $N_{all}$ were the numbers of selected features and
157    the entire features, respectively. $\alpha$ and $\beta$ were parameters whose sum is 1. The larger $\alpha$ is, the
158    more features will be chosen; otherwise, fewer features will be chosen. When the specified
159    number of iterations is reached, the PSO program terminates, and the final selected features will
160    be available.

161
$$cost = \alpha V_{error} + \beta \frac{N_{selected}}{N_{all}} \#(8)$$

162

163 **The validation method**
164 Here, we respectively compared the classification accuracy of the MMPSO method with the
165 results of other algorithms, including mRMR(Peng et al. 2005), ILFS(Roffo et al. 2017),
166 ReliefF(Sui-Yu et al. 2010), Mutinffs(Hutter 2002), FSV(Bradley & Mangasarian 1999),
167 Fisher(Gu et al. 2012), CFS(Hong et al. 2011), UFSOL(Guo et al. 2017), to demonstrate that our
168 method has better classification accuracy. LIBSVM(Chih-Chung & Chih-Jen 2011) is an
169 integrated library, which supports multi-class classification. Here, we performed classification
170 using LIBSVM to test the accuracy with k-fold cross validation.
171 **Summary of datasets**
172 A total of eleven datasets were used in this study; the basic information about the datasets was
173 shown in Table 1. The UCI Machine Learning Repository is a collection of databases, domain
174 theories, and data generators that are used by the machine learning community for the empirical
175 analysis of machine learning algorithms(Dua & Graff 2017). In the beginning, ten datasets that
176 have different numbers of features, instances and classes were downloaded from the UCI website,
177 and they were used to evaluate the performance of our proposed method. Furthermore, we
178 conducted a more thorough analysis to demonstrate the biological application of our method.
179 Tremendous amount of RNA expression data has been produced by large consortium projects
180 such as TCGA and GTEx, creating new opportunities for data mining and deeper understanding
181 of gene functions(Tang et al. 2017). Thus, the final liver hepatocellular carcinoma (LIHC)
182 dataset was obtained from UCSC Xean, which contains large-scale standardized public,
183 multiomic and clinical/phenotype information(Goldman et al. 2020). The LIHC dataset used in
184 the current study contains RNA expression data of over 60,000 genes in 531 biosamples (371
185 tumor samples and 160 normal samples, and the latter further containing 50 normal samples
186 from the TCGA-LIHC cohort and 110 normal tissues from GTEx), and it is available at
187 https://toil-xena-hub.s3.us-east-1.amazonaws.com/download/TCGA-GTEx-TARGET-gene-exp-
188 counts.deseq2-normalized.log2.gz.

189

190 # RESULTS

191 In this section, we focused on testing the accuracy of our proposed MMPSO method and
192 compared it with other methods of mRMR, ILFS, ReliefF, Mutinffs, FSV, Fisher, CFS and
193 UFSOL. All experiments in this study were carried out on a Windows 10 system with an Intel(R)
194 Xeon(R) CPU E5-2420, 1.9 Ghz processor with 16 GB RAM. Our proposed algorithm was
195 implemented in MATLAB 2020b, and the PSO parameters were as follows: population size: 100;
196 number of iterations: 50; $c_1 : 2$; $c_2 : 2$; $\omega : 0.9$; $\alpha : 0.95$; $\beta : 0.05$.
197 **Results of the experiment based on ten datasets**
198 Figure 1 and Table 2 summarized the classification accuracy on the basis of the MMPSO and the
199 compared methods. Here, we defined the threshold $K$ was 100. When the number of original
200 features was greater than 100, the top 100 features from the ranking result were selected as the
201 PSO input for the MMPSO method; otherwise, all features were selected as the input. In Figure 1,
202 we obtained the conclusion that our method was superior to the other methods in terms of
203 classification accuracy based on six datasets including gisette, hillvalley, isolet, madelon, scene
204 and usps. And in the other four datasets, the MMPSO method achieved the accuracies of top

205     three rankings. Therefore, the MMPSO algorithm was outperformed other methods with respect
206     to accuracy of classification by utilizing the same number of features.
207
208     **Results of the experiment based on the biological dataset**
209     After analyzing the first ten datasets, we employed the MMPSO method on the LIHC dataset to
210     identify features (gene biomarkers) that can be used to distinguish the tumor group from the
211     normal group with high accuracy. Different from the previous datasets, LIHC is an unbalanced
212     dataset containing 371 tumor samples and 160 normal samples. Therefore, the preprocessing was
213     needed and the number of genes was reduced from over 60,000 to 15,185 after preprocessing.
214     When the genes were ranked by the mRMR based on MIC method, the top 100 genes were
215     selected and input into PSO to identify the best gene signatures. Finally, we obtained a signature
216     of 18 genes through the MMPSO algorithm, that had better classification compared to other
217     methods. The 18 genes were ACTN1, CACHD1, ERAP2, FAM171A1, FKBP1B, HIST1H2BC,
218     PLVAP, PRKD1, RPL7AP6, ADRA2B, DMKN, FNDC4, NPC1L1, POMC, PYROXD2, RBP1,
219     TRIM29, and ZBED9, with the relative expression of the first nine genes significantly increasing
220     in tumors and the last nine genes decreasing (P < 0.01 in the Wilcoxon rank sum test with
221     continuity correction, Figure 2). Principal component analysis (PCA) was then performed using
222     the 'FactoMineR' and 'Factoextra' packages in R version 4.0.2 based on the expression profiles
223     of the candidate 18 genes. As shown in Figure 3A-B, Dim 1 and Dim 2 were 15% and 11.9%,
224     respectively. Figure 3C illustrated the heatmap of all samples based on the 18 gene expression
225     profiling. The results revealed that the 18-gene signature obtained from MMPSO algorithm
226     could effectively separate the 531 samples into two groups. Since logistic regression show that
227     seven biomarkers of ADRA2B, ERAP2, NPC1L1, PLVAP, POMC, PYROXD2 and TRIM29
228     were significantly associated with *Wald* and *P value*, as shown in Table 3. We further
229     investigated the combined diagnostic efficacy of the seven candidate genes according to the
230     Equation 9.

$$PP = \frac{1}{1 + e^{-(constant + \sum_{1}^{n} coefficient_i * expression_i)}} \#(9)$$

231

232     In the above Equation, *PP* was the functional formula for predicting the incidence of LIHC,
233     i.e. PP-value, and the *constant* and *coefficient* were the result of logistic regression in Table 3.
234     The results of receiver operating characteristic (ROC) curve analysis using MedCalc software
235     based on the value of *PP* and classification labels of LIHC dataset was shown in the Figure 4,
236     which had an area under the curve (AUC) greater than 0.999 and P < 0.0001. It demonstrated
237     that the seven genes significantly distinguished tumors from normal samples in LIHC dataset.
238     To explore whether the 18 genes are associated with survival time of phenotype information in
239     LIHC dataset, the Kaplan-Meier (KM) survival curve was performed by using the "survival" and
240     "survminer" packages in R. For each gene, the cut-off points obtaining from "survminer"
241     package then divided gene expression values into the high (high) and the low (low) groups. We
242     identified that higher expression levels of CACHD1, FKBP1B, PRKD1, and RPL7AP6 were

243  associated with worse overall survival (OS) time, whereas higher expression levels of ADRA2B,
244  PYROXD2 were associated with better OS, as shown in Figure 5.
245
246  **DISCUSSION**
247  High-dimensional data such as text data, multimedia data, aerospace collection data and
248  biometric data have become more common in recent years(Li et al. 2020; Saeys et al. 2007;
249  Wainwright 2019). The need for efficient processing technology for high-dimensional data has
250  become more urgent and challenging. Feature selection, as one of the most popular methods for
251  dimension reduction, plays an important role in high-dimensional data processing, particularly in
252  biological information data(Nguyen et al. 2020; Xue et al. 2015).
253   Generally, features filtered out of the original high-dimensional dataset have more definite
254  physical meanings, making it more convenient for researchers to carry out subsequent work. The
255  direct benefits of feature selection are that it reduces the burden of follow-up work and improves
256  model generalization. Choosing the best subset from the original features has been shown to be a
257  NP-hard problem. When the number of features is $N$, $2^N$ combinations of features must be tried
258  using the greedy strategy, which is unsustainable for ordinary computer systems, especially when
259  the number of features is very large(Faris et al. 2018; Xue et al. 2015). As a result, over the last
260  few decades, some heuristic algorithms have been proposed to find the best subset that can best
261  represent the feature meanings of the original dataset. A best subset can be used to represent the
262  original dataset with the least amount of redundancy between features and the highest correlation
263  between the subset's features and labels. The mRMR algorithm, as an implementation of this
264  mind, can obtain the top $K$ ranking features, where the $K$ value must be manually set and mutual
265  information is used to measure the relevance of two features. The mRMR algorithm is
266  undoubtedly an excellent feature selection framework, and it has been widely used in a variety of
267  fields. Despite this, there are still some shortcomings that can be addressed. On the one hand,
268  mutual information can only handle discrete data, which means that continuous data must be
269  discretized in advance, resulting in some accuracy loss. The output of the mRMR, on the other
270  hand, is the top $K$ features, and there is no "golden rule" to specify a suggested or best $K$ value.
271  To address the above two issues, we proposed a hybrid method called MMPSO. First, the noisy
272  data were removed using a conventional method, and the imbalanced data were corrected using
273  random oversampling technology for preprocessing. Second, we used the MIC(Cao et al. 2021;
274  Reshef et al. 2011) instead of the MI to obtain a more precise correlation value. Although
275  study(Kinney & Atwal 2014) has noted that estimates of mutual information are more equitable
276  than estimates of MIC, there is no denying that MIC has been widely and conveniently used.
277  Furthermore, rapidMic(Tang et al. 2014), an algorithm that can use multiple threads
278  simultaneously, was used to reduce the time expenditure of MIC algorithm. Finally, we selected
279  the top $K$ features from the second step as the input for the PSO algorithm to find the best subset.
280  In accordance with the preceding thought, we conducted our experiment using ten datasets. On
281  these datasets, the MMPSO method was applied to compared with the other methods, including
282  mRMR, ILFS, ReliefF, Mutinffs, FSV, Fisher, CFS, UFSOL. We applied LIBSVM library to
283  evaluate the performance of the three methods and test the classification accuracy of the selected

284 features. The experimental results of the ten datasets provided evidences that the MMPSO
285 method performed better than other feature selection algorithms, when all the algorithms used
286 the same number of features. It's worth mentioning that the mRMR performed similarly to the
287 MMPSO method in the previous findings. Here, we still highlighted the advantages of MMPSO,
288 since it can autonomously select feature subset on the basis of MIC, which will improve the
289 accuracy of classification.

290     To investigate the efficacy of our method for biological data, we used the LIHC dataset (a
291 dataset of RNA expression in liver hepatocellular carcinoma) for further analysis. After
292 performing the MMPSO algorithm, a signature including 18 genes was identified as significant
293 biomarkers for distinguishing between tumor and normal groups. The PCA and ROC analysis
294 results all confirmed that the biomarkers we selected have great discrimination ability, while six
295 biomarkers were significantly associated with the overall survival of patients. Furthermore, we
296 need to discuss the significance of these biomarkers from a biological perspective. Studies(Wang
297 et al. 2014) identified and evaluated tumor vascular PLVAP as a therapeutic target for treatment
298 of HCC but not in nontumorous liver tissues, and this result may provide some clues for the
299 development of drugs for patients with HCC. FNDC4(Wang et al. 2021) was reported to be an
300 extracellular factor and played important roles in the invasion and metastasis of HCC in that it
301 promoted the invasion and metastasis of HCC partly via the PI3K/Akt signaling pathway. Wang
302 et al. (Wang et al. 2019) discovered that PYROXD2 localizes to the mitochondrial inner
303 membrane/matrix, and it plays important roles in regulating mitochondrial function of HCC.
304 TRIM29 plays critical role in many neoplasms. The study(Xu et al. 2018) revealed that higher
305 TRIM29 expression was associated with higher differentiation grade of HCC and its depletion
306 promoted liver cancer cell proliferation, clone formation, migration and invasion. The regulatory
307 role of ACTNs in tumorigenesis has been demonstrated and ACTN1 was significantly
308 upregulated in HCC tissue and closely related to tumor size, TNM stage and patient
309 prognoses(Chen et al. 2021b). Based on the above studies, the majority of the genes identified by
310 our algorithm are promising candidate biomarkers for the diagnosis or treatment of liver cancer.
311

## CONCLUSION

313 In this paper, we proposed the MMPSO hybrid algorithm to identify a feature subset for high-
314 dimensional dataset. The experimental data provided evidences that our method outperformed
315 others. More importantly, by applying our proposed algorithm to the biological LIHC dataset, we
316 obtained the gene signatures in classifying tumors and normal samples with high efficacy. Our
317 study also has several limitations. Despite the fact that we used rapidMic to accelerate the
318 calculation, the computational complexity for too many features remains relatively high. In
319 addition, we selected only the top $K$ features as the PSO input without a theoretical foundation.
320 Therefore, there is still some space for improvement in selecting a better subset of features for
321 high-dimensional datasets, and we will advance this in future works.
322

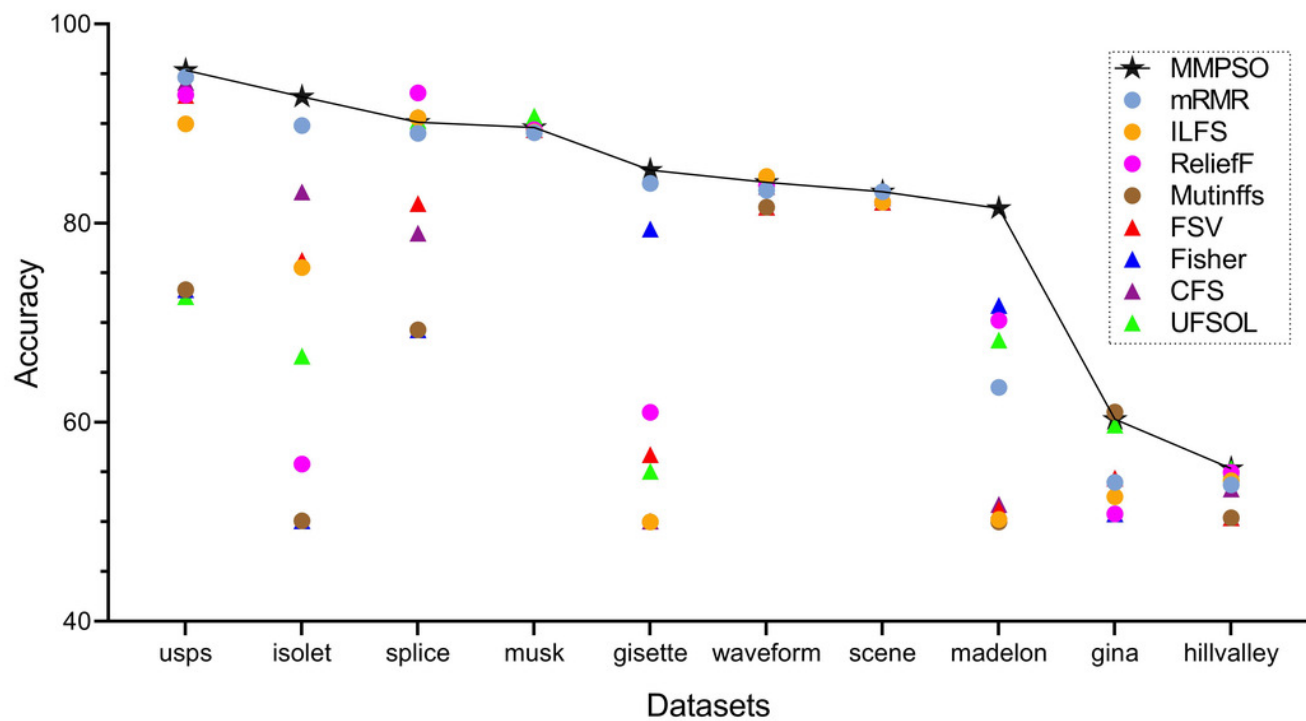
## References

324    Bradley PS, and Mangasarian OLJMKPI. 1999. Feature Selection via Concave Minimization
325          and Support Vector Machines.
326    Cao D, Chen Y, Chen J, Zhang H, and Yuan Z. 2021. An improved algorithm for the maximal
327          information coefficient and its application. *Royal Society Open Science* 8.
328          10.1098/rsos.201424
329    Chen K, Xue B, Zhang M, and Zhou F. 2021a. An Evolutionary Multitasking-Based Feature
330          Selection Method for High-Dimensional Classification. *IEEE transactions on cybernetics*
331          PP.
332    Chen Q, Zhou X-W, Zhang A-J, He KJJoE, and Research CC. 2021b. ACTN1 supports tumor
333          growth by inhibiting Hippo signaling in hepatocellular carcinoma.  40:1-13.
334    Chen Z, Pang M, Zhao Z, Li S, Miao R, Zhang Y, Feng X, Feng X, Zhang Y, Duan M, Huang L,
335          and Zhou F. 2020. Feature selection may improve deep neural networks for the
336          bioinformatics problems. *Bioinformatics* 36:1542-1552. 10.1093/bioinformatics/btz763
337    Chih-Chung C, and Chih-Jen L. 2011. Libsvm: a library for support vector machines.
338    Chuang LY, Chang HW, Tu CJ, Yang CHJCB, and Chemistry. 2008. Improved binary PSO for
339          feature selection using gene expression data.  32:29-38.
340    Ding C, and Peng H. 2005. Minimum redundancy feature selection from microarray gene
341          expression data. *Journal of Bioinformatics \& Computational Biology* 3:185-205.
342    Dua D, and Graff C. 2017. {UCI} Machine Learning Repository.
343    Eberhart R, and Kennedy J. 2002. A new optimizer using particle swarm theory. Mhs95 Sixth
344          International Symposium on Micro Machine \& Human Science.
345    Faris H, Mafarja MM, Heidari AA, Aljarah I, and Fujita HJK-BS. 2018. An Efficient Binary Salp
346          Swarm Algorithm with Crossover Scheme for Feature Selection Problems.
347    Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, Kamath A, Banerjee A, Luo Y, Rogers
348          D, Brooks AN, Zhu J, and Haussler D. 2020. Visualizing and interpreting cancer
349          genomics data via the Xena platform. *Nature Biotechnology* 38:675-678.
350          10.1038/s41587-020-0546-8
351    Gu Q, Li Z, and Han J. 2012. Generalized Fisher Score for Feature Selection.
352    Guo J, Guo Y, Kong X, and Ran H. 2017. Unsupervised feature selection with ordinal locality.
353          2017 IEEE International Conference on Multimedia and Expo (ICME).
354    Holland JHJaa. 1975. Adaptation in Natural and Artificial Systems.
355    Hong Z, Cheung YMJIToPA, and Intelligence M. 2011. Feature Selection and Kernel Learning
356          for Local Learning-Based Clustering.  33:1532-1547.
357    Hutter M. 2002. Robust Feature Selection using Distributions of Mutual Information.
358    Kinney JB, and Atwal GS. 2014. Equitability, mutual information, and the maximal information
359          coefficient. *Proceedings of the National Academy of Sciences of the United States of
360          America* 111:3354-3359. 10.1073/pnas.1309933111
361    Larranaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armananzas R,
362          Santafe G, Perez A, and Robles V. 2006. Machine learning in bioinformatics. *Briefings in
363          Bioinformatics* 7:86-112. 10.1093/bib/bbk007
364    Li B, Wang L, and Song W. 2008a. Ant Colony Optimization for the Traveling Salesman
365          Problem Based on Ants with Memory.496-+. 10.1109/icnc.2008.354
366    Li BQ, Tao H, Lei L, Cai YD, Kuo-Chen C, and Lee HPJPONE. 2012. Identification of Colorectal
367          Cancer Related Genes with mRMR and Shortest Path in Protein-Protein Interaction
368          Network.  7:e33393-.
369    Li G, Hu X, Shen X, Chen X, and Li Z. 2008b. A novel unsupervised feature selection method
370          for bioinformatics data sets through feature clustering. *2008 IEEE International
371          Conference on Granular Computing, GRC 2008*:41-47. 10.1109/GRC.2008.4664788
372    Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, and Liu HJACS. 2016. Feature
373          Selection: A Data Perspective.  50.

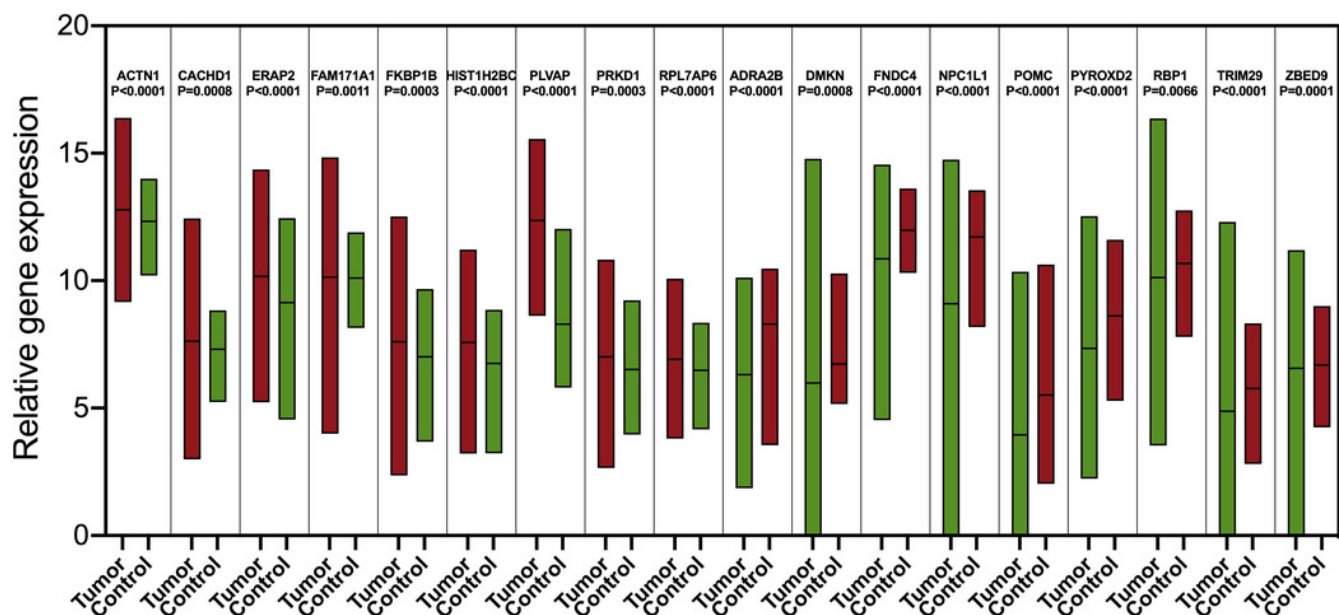
374  Li X, Wu C, and Li PJPotACoAI. 2020. IVFS: Simple and Efficient Feature Selection for High
375      Dimensional Topology Preservation. 34:4747-4754.
376  Liu J, Liu Y, and Zhang Q. 2022. A Weight Initialization Method Based on Neural Network with
377      Asymmetric Activation Function. *Neurocomputing*.
378  Liu Y, Wang G, Chen H, Dong H, Zhu X, and Wang S. 2011. An improved particle swarm
379      optimization for feature selection. *Journal of Bionic Engineering* 8:191-200.
380  Ma A, Re B, Hm A, and Hn AJG. 2020. Heuristic filter feature selection methods for medical
381      datasets. 112:1173-1181.
382  Mundra PA, and Rajapakse JCJIToN. 2010. SVM-RFE With MRMR Filter for Gene Selection.
383      9:31-37.
384  Nguyen BH, Xue B, and Zhang M. 2020. A survey on swarm intelligence approaches to feature
385      selection in data mining. *Swarm and Evolutionary Computation* 54:100663.
386  Obayashi T, and Kinoshita K. 2009. Rank of correlation coefficient as a comparable measure for
387      biological significance of gene coexpression. *DNA research* 16:249-260.
388  Peng HC, Long FH, and Ding C. 2005. Feature selection based on mutual information: Criteria
389      of max-dependency, max-relevance, and min-redundancy. *Ieee Transactions on Pattern
390      Analysis and Machine Intelligence* 27:1226-1238. 10.1109/tpami.2005.159
391  Rasheed AAJTJoC, and Education M. 2021. Feature Selection: An Assessment of Some
392      Evolving Methodologies. 12:1982-1988.
393  Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Lander ES,
394      Mitzenmacher M, and Sabeti PC. 2011. Detecting Novel Associations in Large Data Sets
395      %J Science. 334.
396  Robnik-Šikonja M, and Kononenko I. 2003. Theoretical and Empirical Analysis of ReliefF and
397      RReliefF %J Machine Learning. 53.
398  Roffo G, Melzi S, Castellani U, and Vinciarelli AJI. 2017. Infinite Latent Feature Selection: A
399      Probabilistic Latent Graph-Based Ranking Approach.
400  Saeys Y, Inza I, and Larrañaga P. 2007. A review of feature selection techniques in
401      bioinformatics. *Bioinformatics* 23:2507-2517. 10.1093/bioinformatics/btm344
402  Shreem SS, Abdullah S, Nazri M, Alzaqebah MJJoT, and Technology AI. 2013. Hybridizing
403      relief, mRMR filters and GA wrapper approaches for gene selection. 46:1034-1039.
404  Stefano CD, Fontanella F, and Freca A. 2017. Feature Selection in High Dimensional Data by a
405      Filter-Based Genetic Algorithm. European Conference on the Applications of
406      Evolutionary Computation.
407  Sui-Yu, Analysis WJP, and Applications. 2010. Huan Liu and Hiroshi Motoda (eds):
408      Computational methods of feature selection.
409  Tang D, Wang M, Zheng W, and Wang H. 2014. RapidMic: Rapid Computation of the Maximal
410      Information Coefficient. *Evolutionary bioinformatics online* 10:11-16.
411      10.4137/EBO.S13121
412  Tang Z, Li C, Kang B, Gao G, Li C, and Zhang Z. 2017. GEPIA: a web server for cancer and
413      normal gene expression profiling and interactive analyses. *Nucleic Acids Res* 45:W98-
414      W102. 10.1093/nar/gkx247
415  Tao H, Hou C, Nie F, Jiao Y, and Yi D. 2015. Effective Discriminative Feature Selection with
416      Non-trivial Solutions.
417  Van Hulse J, Khoshgoftaar TM, and Napolitano A. 2011. A comparative evaluation of feature
418      ranking methods for high dimensional bioinformatics data. *Proceedings of the 2011 IEEE
419      International Conference on Information Reuse and Integration, IRI 2011*:315-320.
420      10.1109/IRI.2011.6009566
421  Villa A, Mundanad Narayanan A, Van Huffel S, Bertrand A, and Varon C. 2021. Utility metric for
422      unsupervised feature selection. *PeerJ Comput Sci* 7:e477. 10.7717/peerj-cs.477
423  Wainwright MJ. 2019. High-dimensional statistics: A non-asymptotic viewpoint. 48.

424    Wang B, Zheng B, Lu Y, Huang D, Liu J, Song J, and Zheng S. 2021. FNDC4 acts as an
425          extracellular factor to promote the invasiveness of hepatocellular carcinoma partly via
426          the PI3K/Akt signalling pathway. *Cancer Med* 10:7242-7252. 10.1002/cam4.4225
427    Wang L, Wang Y, and Chang Q. 2016. Feature selection methods for big data bioinformatics: A
428          survey from the search perspective. *Methods* 111:21-31. 10.1016/j.ymeth.2016.08.014
429    Wang T, Xie X, Liu H, Chen F, Du J, Wang X, Jiang X, Yu F, and Fan H. 2019. Pyridine
430          nucleotide-disulphide oxidoreductase domain 2 (PYROXD2): Role in mitochondrial
431          function. *Mitochondrion* 47:114-124. 10.1016/j.mito.2019.05.007
432    Wang YH, Cheng TY, Chen TY, Chang KM, Chuang VP, and Kao KJ. 2014. Plasmalemmal
433          Vesicle Associated Protein (PLVAP) as a therapeutic target for treatment of
434          hepatocellular carcinoma. *BMC Cancer* 14:815. 10.1186/1471-2407-14-815
435    Xiangyang, Wang, and, Jie, Yang, and, Xiaolong, Teng, and, and Letters WJPR. 2007. Feature
436          selection based on rough sets and particle swarm optimization.
437    Xu M, Hu J, Zhou B, Zhong Y, Lin N, and Xu R. 2018. TRIM29 prevents hepatocellular
438          carcinoma progression by inhibiting Wnt/β-catenin signaling pathway. *Acta Biochimica et*
439          *Biophysica Sinica* 51:68-77. 10.1093/abbs/gmy151 %J Acta Biochimica et Biophysica
440          Sinica
441    Xue B, Zhang M, and Browne WN. 2012. New fitness functions in binary particle swarm
442          optimisation for feature selection. 2012 IEEE congress on evolutionary computation. p 1-
443          8.
444    Xue B, Zhang M, Browne WN, and Yao X. 2015. A survey on evolutionary computation
445          approaches to feature selection. *Ieee Transactions on Evolutionary Computation* 20:606-
446          626.
447    Yuanyuan H, Lan H, and Fengfeng Z. 2021. A dynamic recursive feature elimination framework
448          (dRFE) to further refine a set of OMIC biomarkers %J Bioinformatics.  37.
449

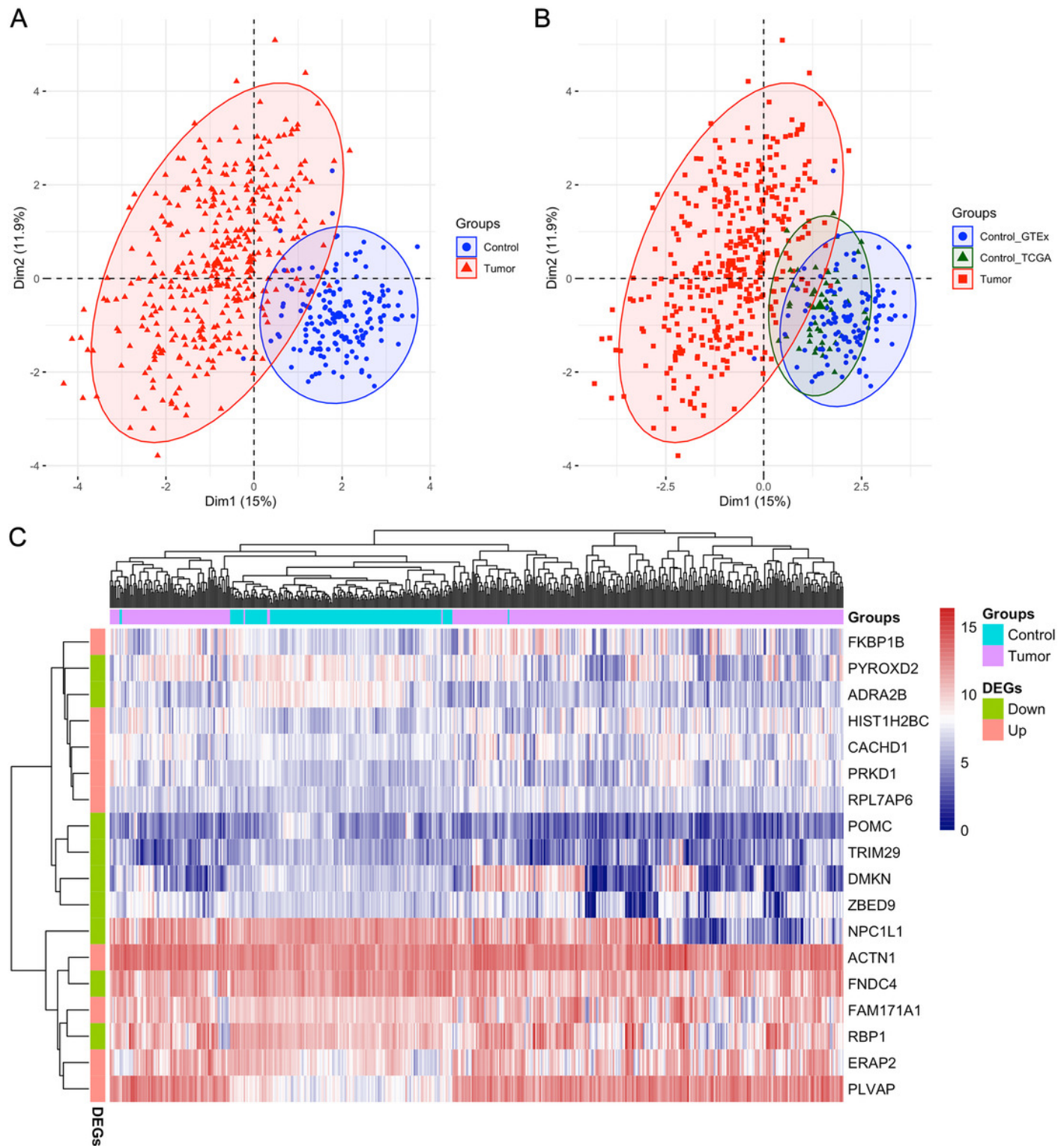
# Figure 1

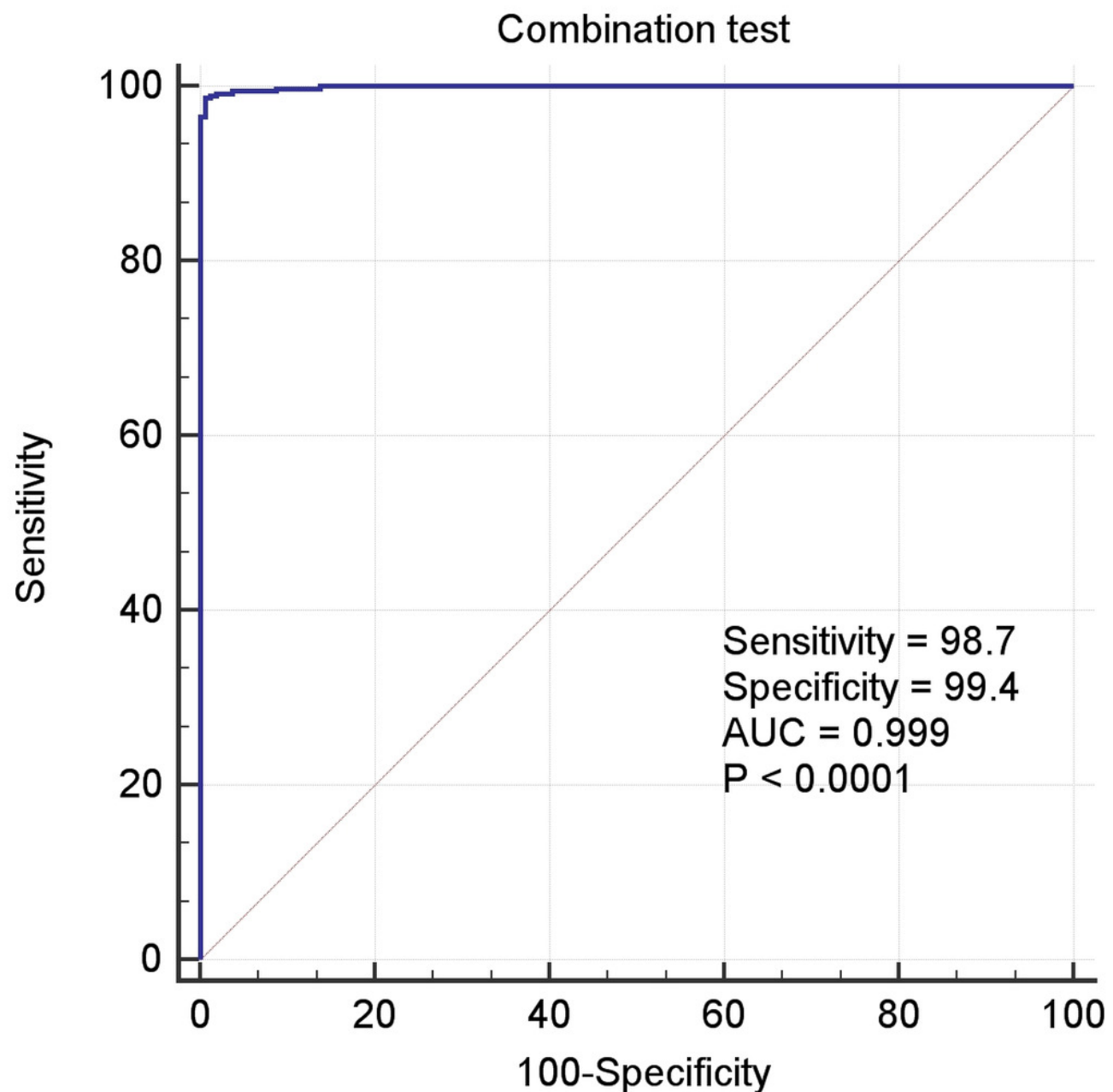
The accuracy of nine algorithms based on ten datasets

# Figure 2

The relative expressions of the 18 genes in tumor and control groups.

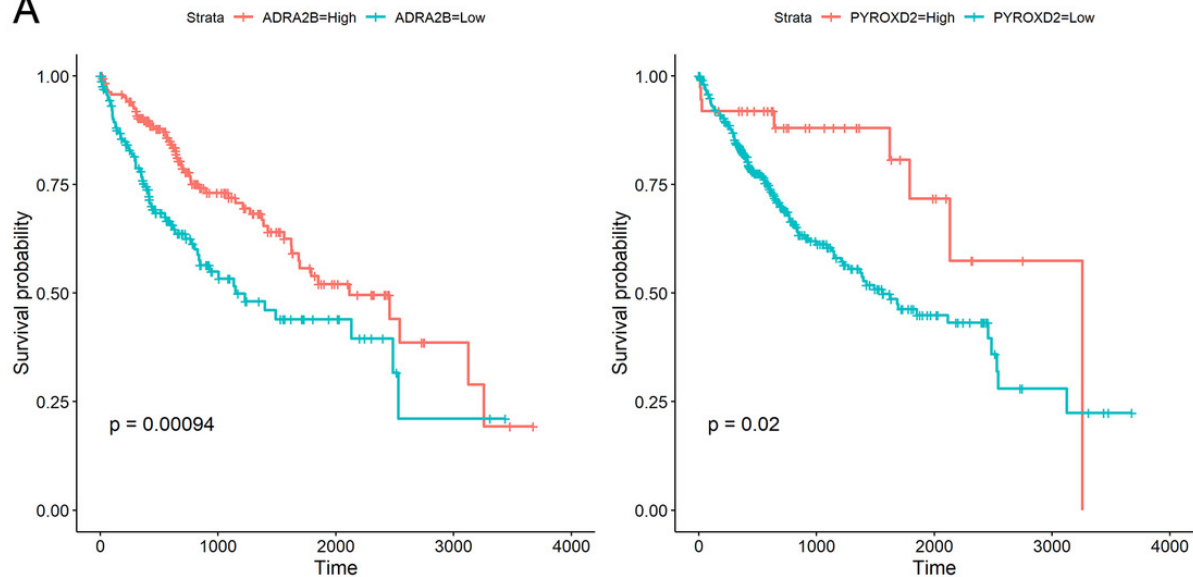
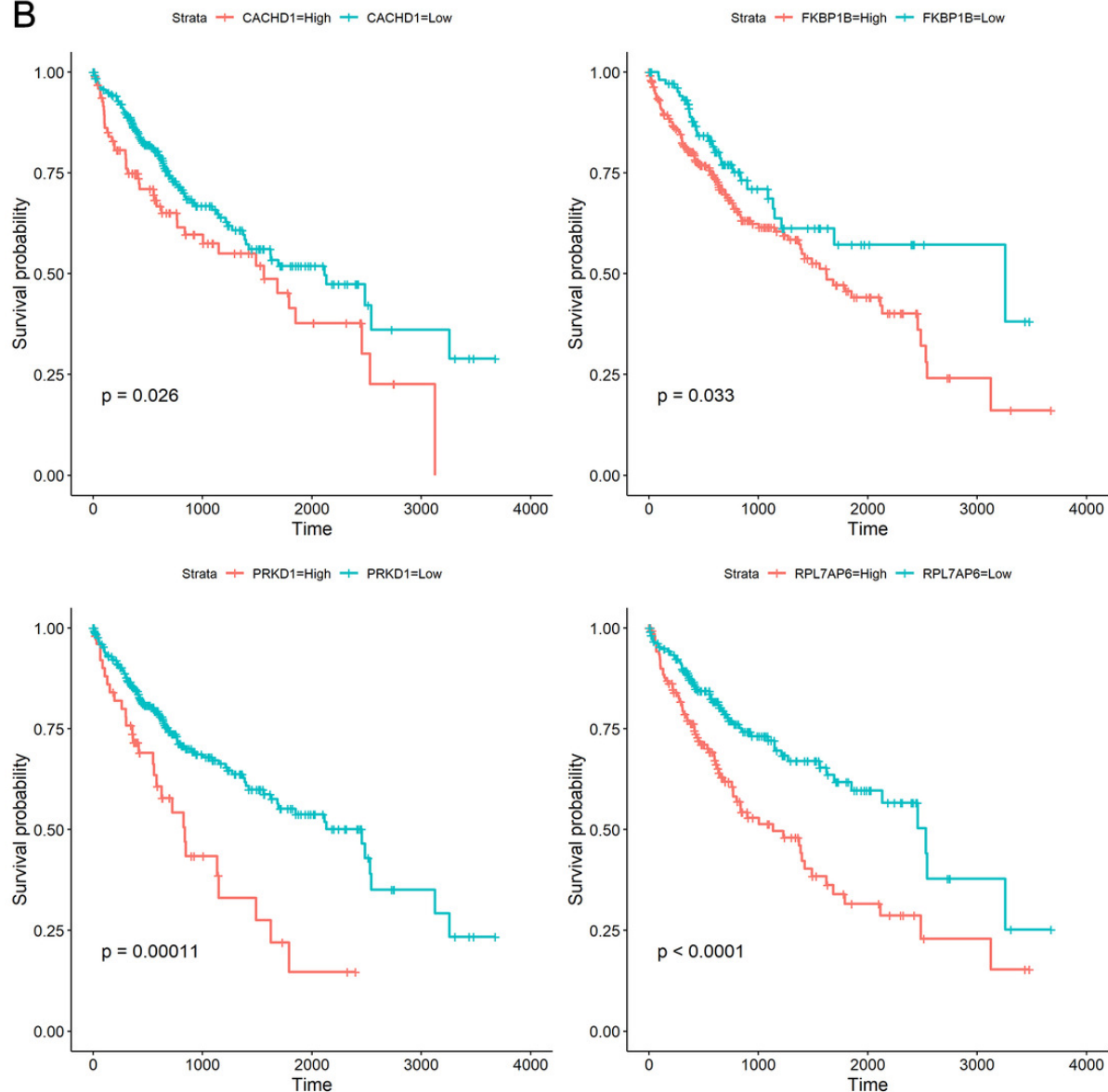
The values were displayed as floating bars (min to max) with a line at the mean value. The first nine genes increased in tumors (color in red) compared to controls (color in green), while the last nine genes decreased in tumors (color in green) compared to controls (color in red). The statistic was performed by Wilcoxon rank sum test with continuity correction in R.

# Figure 3

PCA and heatmap analysis of the 18 gene signatures obtained from MMPSO method in LIHC dataset.

(A) PCA analysis of tumor and control samples. The PCA analysis was performed by using "FactoMineR" and "factoextra" packages in R; (B) PCA analysis of tumor and control samples, the latter including control_GTEx and control_TCGA samples; (C) Heatmap of all the samples based on the 18 gene expression profiling. The heatmap analysis was performed by using "pheatmap" package in R. DEGs: differentially expressed genes. The expression levels of up DEGs were increased in tumors compared to controls, and the down DEGs were decreased.

# Figure 4

The ROC curves of the combination of seven candidate biomarkers ( ADRA2B, ERAP2, NPC1L1, PLVAP, POMC, PYROXD2 and TRIM29).

# Figure 5

Kaplan-Meier curves for prognostic analysis of six genes, including ADRA2B, PYROXD2, CACHD1, FKBP1B, PRKD1 and RPL7AP6.

(A) The Kaplan-Meier curves based on two down-regulated genes (ADRA2B, PYROXD2); (B) The Kaplan-Meier curves based on four up-regulated genes (CACHD1, FKBP1B, PRKD1 and RPL7AP6). The horizontal axis represents the survival time (days), and the vertical axis represents the overall survival rate.

Basic information of the datasets in this study.

The ten datasets obtained from UCI were analyzed to demonstrate the superiority of MMPSO; the biological dataset was used as an application of the proposed MMPSO method.

Table 1. Basic information of the datasets in this study

| Datasets | | Instances | Features | Classes |
|---|---|---|---|---|
| Datasets obtained from UCI | gina | 3468 | 970 | 2 |
| | gisette | 6000 | 5000 | 2 |
| | hillvalley | 1212 | 100 | 2 |
| | isolet | 7797 | 617 | 26 |
| | madelon | 2000 | 500 | 2 |
| | musk | 6598 | 166 | 2 |
| | scene | 2407 | 294 | 2 |
| | splice | 3190 | 60 | 3 |
| | usps | 9298 | 256 | 10 |
| | waveform | 5000 | 21 | 3 |
| Biological dataset | LIHC | 531 | 60498 | 2 |

2   The ten datasets obtained from UCI were analyzed to demonstrate the superiority of MMPSO; the

3   biological dataset was used as an application of the proposed MMPSO method.

Accuracy of algorithms based on the ten datasets .

1                              Table 2. Accuracy of algorithms based on the ten datasets.

| Datasets | MMPSO | mRMR | ILFS | ReliefF | Mutinffs | FSV | Fisher | CFS | UFSOL |
|---|---|---|---|---|---|---|---|---|---|
| gina | 60.32 | 53.97 | 52.53 | 50.79 | **61.04** | 54.40 | 50.79 | 50.79 | 59.74 |
| gisette | **85.33** | 84.00 | 50.00 | 61.00 | 50.00 | 56.75 | 79.42 | 50.08 | 55.08 |
| hillvalley | **55.37** | 53.72 | 54.13 | 54.96 | 50.41 | 50.41 | 50.41 | 53.31 | 55.37 |
| isolet | **92.69** | 89.80 | 75.56 | 55.81 | 50.10 | 76.27 | 50.10 | 83.13 | 66.65 |
| madelon | **81.50** | 63.50 | 50.25 | 70.25 | 50.00 | 51.25 | 71.75 | 51.75 | 68.25 |
| musk | 89.61 | 89.08 | 89.16 | 89.39 | 89.39 | 89.39 | 89.39 | 89.39 | **90.75** |
| scene | **83.16** | 83.16 | 82.12 | 82.12 | 82.12 | 82.12 | 82.12 | 82.54 | 82.12 |
| splice | 90.13 | 89.02 | 90.60 | **93.10** | 69.28 | 81.97 | 69.28 | 79.00 | 90.28 |
| usps | **95.37** | 94.67 | 89.99 | 92.90 | 73.32 | 92.85 | 73.32 | 94.14 | 72.62 |
| waveform | 84.10 | 83.30 | **84.70** | 83.80 | 81.60 | 81.60 | 81.60 | 83.70 | 83.70 |

Logistic regression analysis of the independent significance of the 18 genes as diagnostic biomarkers.

1  Table 3. Logistic regression analysis of the independent significance of the 18 genes as
2  diagnostic biomarkers.

| Variable | Coefficient | Std. Error | Wald | *P* value | Odds ratio | 95% CI |
|---|---|---|---|---|---|---|
| ADRA2B | -1.2426 | 0.46552 | 7.1249 | 0.0076 | 0.2886 | 0.1159 to 0.7188 |
| ERAP2 | 0.74916 | 0.33879 | 4.8897 | 0.027 | 2.1152 | 1.0889 to 4.1090 |
| NPC1L1 | -0.72899 | 0.34876 | 4.3691 | 0.0366 | 0.4824 | 0.2435 to 0.9556 |
| PLVAP | 3.42399 | 0.83386 | 16.8608 | <0.0001 | 30.6915 | 5.9872 to 157.331 |
| POMC | -0.80808 | 0.32528 | 6.1715 | 0.013 | 0.4457 | 0.2356 to 0.8432 |
| PYROXD2 | -0.64701 | 0.32983 | 3.848 | 0.0498 | 0.5236 | 0.2743 to 0.9995 |
| TRIM29 | -0.99483 | 0.49833 | 3.9853 | 0.0459 | 0.3698 | 0.1392 to 0.9821 |
| Constant | 1.81243 | 8.70136 | 0.04339 | 0.8350 | | |