# Feature Selection for Optimized High-Dimensional Biomedical Data Using an Improved Shuffled Frog Leaping Algorithm

Bin Hu, Yongqiang Dai, Yun Su, Philip Moore, Xiaowei Zhang, Chengsheng Mao, Jing Chen, and Lixin Xu

**Abstract**—High dimensional biomedical datasets contain thousands of features which can be used in molecular diagnosis of disease, however, such datasets contain many irrelevant or weak correlation features which influence the predictive accuracy of diagnosis. Without a feature selection algorithm, it is difficult for the existing classification techniques to accurately identify patterns in the features. The purpose of feature selection is to not only identify a feature subset from an original set of features [without reducing the predictive accuracy of classification algorithm] but also reduce the computation overhead in data mining. In this paper, we present our improved shuffled frog leaping algorithm which introduces a chaos memory weight factor, an absolute balance group strategy, and an adaptive transfer factor. Our proposed approach explores the space of possible subsets to obtain the set of features that maximizes the predictive accuracy and minimizes irrelevant features in high-dimensional biomedical data. To evaluate the effectiveness of our proposed method, we have employed the K-nearest neighbor method with a comparative analysis in which we compare our proposed approach with genetic algorithms, particle swarm optimization, and the shuffled frog leaping algorithm. Experimental results show that our improved algorithm achieves improvements in the identification of relevant subsets and in classification accuracy.

**Index Terms**—Shuffled frog leaping algorithm, feature selection, k-nearest neighbor, classification accuracy, biomedical data

## 1 Introduction

The analysis of disease data is a very important in biomedicine and bioinformatics, however, while treatment of diseases such as cancer has achieved impressive results there remain conditions for which effective treatment has yet to be achieved. Early diagnosis plays an important role for clinicians and patients in the control and management of disease [1] and high dimensional biomedical data sets have been used in diagnosis. However such datasets contain a large number of irrelevant or weak correlation features [2] and for existing classification techniques it is difficult to accurately identify the patterns from these features [3].

Feature selection may be viewed in terms of the identification and selection of a subset of features from an original set of features forming patterns in a given dataset. The subset should be ‘necessary and sufficient’ to describe target concepts while retaining suitably high accuracy in the representation of the original features. The selection of relevant features [with the elimination of irrelevant features] can: (1) improve classification accuracy, and (2) reduce the learning period.

In ’real-world’ applications, high-dimensional biomedical data sets are generally very large and contain tens of thousands of features where there are in total 2 competing candidate subsets for any (*F*) number of features. There may be many features in biomedical data sets that ere either irrelevant or exhibit a weak correlation, the identification of the optimal feature subset may not only eliminate redundant information but also reduce the computational cost required in data mining while improving classification accuracy. The effective identification and selection of relevant candidate subsets requires an efficient and effective search method and learning algorithm; however, the development of such methods and learning algorithms designed to identify optimal subsets remains an open research question. In this paper we present an approach to enable feature selection from high-dimensional biomedical data based on the *Shuffled Frog Leaping Algorithm* (SFLA).

The SFLA is one of a number of nature inspired algorithms based on the swarm intelligence [4]. Because of its characteristics, which include: (1) a simple concept, (2) reduced parameters, (3) powerful optimal performance, (4) fast calculation speed, and (5) easy realization, it has been applied in many fields including model identification Problems [5], [6], scheduling problems [7], [8], parameter optimization problems [9], the traveling salesman problem [10], the unit commitment problem [11], the distribution problem [12] and the controller problem [13]. However, our literature search has failed to identify documented research related to feature selection using the SFLA.

In this paper, we propose an approach which introduces an extension to the SLFA to develop the *Improved Shuffled Frog Leaping Algorithm* (ISFLA); the ISFLA introduces [to the updating strategy] a *chaos memory weight factor* (CMWF), an

- *B. Hu, Y. Su, P. Moore, X. Zhang, C. Mao, J. Chen, and L. Xu are with the School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China. E-mail: {bh, suy13, zhangxw, maocs11, chenj12, xulx13}@lzu.edu.cn, ptmbcu@gmail.com.*
- *Y. Dai is with the School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China, and the College of Information Science and Technology, Gansu Agricultural University, Lanzhou 730070, China. E-mail: daiyq14@lzu.edu.cn.*

*Manuscript received 25 Feb. 2016; revised 15 June 2016; accepted 15 Aug. 2016. Date of publication 24 Aug. 2016; date of current version 6 Dec. 2018.*
*For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.*
*Digital Object Identifier no. 10.1109/TCBB.2016.2602263*

*absolute balance group strategy* (ABGS) and an *adaptive transfer factor* (ATF). The ISFLA is discussed in detail in subsequent sections of this paper. We have employed the K-nearest neighbor method with high-dimensional biomedical data to evaluate the ISFLA including carrying out a comparative analysis with an *improvement Genetic Algorithm* (IGA), *improvement Particle Swarm Optimization* (IPSO) and the SFLA. Experimental results show that our ISFLA demonstrates improved performance in respect of the identification of relevant subsets with greater classification accuracy as compared to the alternative methods.

This paper is structured as follows: related research is considered in Section 2 with the SFLA introduced in Section 3. The extended ISFLA is presented in Section 4 with the evaluation function discussed in Section 5. In Section 6 we set out our experimental results with a discussion. The paper closes with Section 7 where we provide concluding observations.

## 2 Related Research

There are a number of feature selection algorithms documented in the literature [14]. Feature selection, a combinatorial optimization problem and an intelligent optimization algorithm [15], has been proposed as an effective solution to the feature selection problem. To address the challenges in computer supported diagnosis of skin tumors [in dermatology], an approach using GA has been proposed in [15] to enable feature selection including the capability to identify feature subsets for the recognition process with improvements in classification performance.

A GA has also been employed in [16] where a GA is combined with the nearest neighbor method for the discrimination of seeds by artificial vision. The value to be maximized was the percentage of correct classification by the 'leave-one-out nearest neighbour' method. The best results were obtained with an initialisation probability of 0.1, at generation 400, leading to a 3.00 percent omisclassification between the four seed species.

Zhang et al in [17] present a binary "feature selection algorithm based on bare bones particle swarm optimization" which has been successfully applied to solve feature selection problems. Quantitative results demonstrate that the approach proposed achieves the best average classification accuracies for seven out of eight data sets used in the evaluation.

Yang et al [18] propose an approach to feature Selection using *Memetic Algorithms* (MA). Yang et al claim that their proposed approach is capable of producing high classification accuracy with a small number of features and is superior to GA and PSO methods in terms of accuracy, particularly for large-sized problems. The reported experimental results show that this method enables increased efficiency in search and is capable of producing high classification accuracy with a small number of features.

Kabir et al [19] present a new hybrid *Ant Colony Optimization Algorithm for Feature Selection* (ACOFS). It is claimed that this approach not only provides an effective balance between exploration and exploitation of ants in the search, but also intensifies the global search capability of ant colony optimization in the realization of high quality solutions in feature selection problems. Reported results in experimental testing show that ACOFS provides the remarkable ability to generate reduced-size subsets [of salient features] while yielding significant classification accuracy.

Huang et al in [20] also approach the problem of feature selection using a nature inspired solution, the proposed approach employs a new hybrid approach based on *Particle Swarm Optimization and Support Vector Machines* (PSO-SVM) with feature selection and parameter optimization to solve feature subset selection with the setting of kernel parameters. A data mining system, implemented via a distributed architecture using web service technology, is used to reduce the computational time. The experimental results show that the proposed approach can correctly select the discriminating input features while achieving high classification accuracy.

Wang et al in [21] introduce a novel *Ant Colony Optimization* (ACO) method to enable feature selection based on rough sets and PSO to classify *Hand Motion Surface Electromyography* (SEMG) signals. In a comparative analysis using *principal component analysis* (PCA), experimental results demonstrate that the proposed method [based feature selection] can achieve high classification rates in SEMG motion classification task.

An *Improved Binary Particle Swarm Optimization* (IBPSO) approach is proposed in [22] to implement feature selection, a K-nearest neighbor (K-NN) serving as an evaluator of IBPSO for gene expression data classification problems. Experimental results show the method effectively simplified gene (feature) selection and reduced the total number of genes (features) required.

The related research considered has demonstrated that nature inspired systems represent an effective basis upon which feature selection may be achieved. In this paper we have applied a nature inspired approach using our novel extended SFLA [the ISFLA] for high-dimensional biomedical data feature selection.

## 3 Shuffled Frog Leaping Algorithm

In the SFLA, according to the fitness [from big/small to small/big], the individuals are assigned to several groups in turn where the worst individual ($P_w^t$) has learned from the best individual ($P_b^t$) in a subgroup. If there is no progress, ($P_w^t$) will learn from the global best individual ($P_g^t$). If there is still no progress, ($P_w^t$) will be replaced by a random individual. The number of iterations in the algorithm is given by (***t***).

$$Dis^t = R \times (P_b^t - P_w^t) \tag{1}$$

$$P_w^{t+1} = P_w^t + Dis^t\ (Dis_m \geq Dis \geq -Dis_m), \tag{2}$$

where: (1) $P_w^{t+1} = (P_{w1}^{t+1}, P_{w2}^{t+1}, ..., P_{wn}^{t+1})$ is a new individual generated by the updating strategy, (2) $Dis^t$ is each moving step length. (3) $R$ is a random number with a change range of [0…1], and (4) [-$\mathrm{Dis_m}$, $\mathrm{Dis_m}$] is the values range of the leaping step.

Following updating, if the newly generated ($P_w^{t+1}$) is an improvement over the old ($P_w^t$), ($P_w^t$) will be replace by ($P_w^{t+1}$); otherwise, ($P_b^t$) will be replace by ($P_g^t$). If ($P_w^t$) still shows no progress, it will be replaced randomly by a new individual. This is an iterative process with the number of iterations being equal to the number of subgroup individuals. When the subgroup processing is complete all subgroups will be randomly sorted and re-divided into new subgroups, the process being repeated until the pre-determined termination criteria is satisfied.

In considering the SFLA the literature documents a number of methods designed improve the performance of the algorithm, the performance improvements being generally addressing updating strategies. Luo et al in [23] propose a "*power law extremal optimization neighborhood*" search method designed to improve the speed of search. Wang and Fang in [24] introduce a method in which virtual frogs are encoded as the extended multi-mode activity list and decoded by the multi-mode serial schedule generation scheme; the ISFLA (see Section 4) is proposed as an effective approach to solve the multi mode resource constrained project scheduling problem.

Li et al in [25] propose a modified SFLA which improves the leaping rule by extending the leaping step size and adding a leaping inertia component to account for social behavior, this approach is designed to further improve the local search ability and speed up convergence. An adaptive SFLA [26] has been proposed and applied to economic dispatch problems.

The reported results of the research identified demonstrate the efficacy of the SFLA approach.

## 4 Improved Shuffled Frog Leaping Algorithm

In this paper we propose the ISFLA which is an extension to the SFLA. The ISFLA implements a new strategy to improve the performance. In the following sections we introduce our approach, the evaluation function is discussed in Section 5 with the experimental results and discussion being presented in Section 6.

### 4.1 Improvement Strategy

The proposed strategy implements three components: (1) CMWF, (2) ABGS, and (3) ATF. These components are discussed in the following sections.

*1) CMWF:* The balance between the global exploration and local search ability was controlled by the chaos memory weight. Using a large memory weight, the global search ability was improved, while using a small memory weight the local search ability was improved. Proper modulation of the memory weight value is important.

The memory weight (*w*) is the key factor influencing the convergence and thus will greatly affect the algorithm search process and classification accuracy. The algorithm process often suffers from the potential to identify an individual in a local optimum resulting in premature convergence. To address the potential for early convergence and achieve improved classification results we have employed a chaos memory weight.

Chaos is a deterministic dynamic system that is very sensitive to initial conditions and parameters. The nature of chaos is apparently random and unpredictable, however, it also possesses an element of regularity as discussed in [27]. A *chaotic map* is used to determine the memory weight value in the iteration process. *Logistic Maps* (LM) are generally recognized as the most frequently used *chaotic behavior maps* and LM provide an expressive representation of unstable dynamic behavior.

Chaotic sequences have been shown to provide the capability to implement fast generation and storage as there is no requirement to store long sequences [27]. In this paper, chaotic sequences are introduced in the algorithm with LM determining the memory weight values. The number of iterations in the algorithm is given by (*t*) and the memory weight value is modified by a LM according to

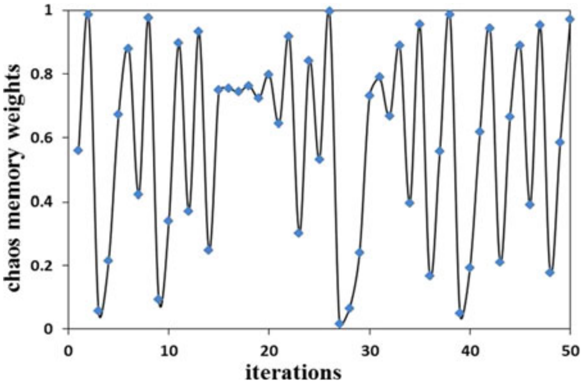


Fig. 1. Chaotic memory weight using logistic map.

$$\begin{gathered}Dis^{t+1} = W(t+1) \times Dis^{t} + R \times (P_b^t - P_w^t)\\ w(t+1) = 4.0 \times w(t) \times (1 - w(t))(Dis_m \geq Dis \geq -Dis_m).\end{gathered} \tag{3}$$

Fig. 1 shows the chaotic memory weight value using a LM for the total number of iterations with the chaotic memory weight value $w(0)$ set to 0.56. When the memory weight value approaches [1] the algorithm strengthens the global search ability. For memory weight values approaches [*0*], the algorithm enhances the local search capability.

*2) ABGS:* In considering the basic SFLA, individuals are sorted and assigned to each group iteratively according to the fitness (fitness value from big/small to small/big). The fittest individuals are assigned to the first group with the worst individuals assigned to the final group. For example consider:

- 100 individuals are assigned to 10 groups, for each group of 10 individuals the 1st, 11th, ... 91st are assigned to the first group with the 2nd, 12th ... 92nd assigned to the second group.
- This process is repeated until the 10th, 20th, ... 100th are assigned to the final group. We call this grouping strategy the *Classic Grouping Strategy* (CGS).
- In each layer, the fittest individuals are concentrated in the first group; in such a situation the number individuals in the first group update ($\boldsymbol{P_g}$) will be greater than other groups.

Once the individuals in the first group are limited to a local optimum, the algorithm may find it difficult to escape from the local optimum and thus avoid premature convergence. In such situation, for the SFLA, there may be a detrimental effect on the classification performance. To reduce the dependency on the first group [and escape from the premature convergence] we must balance the fitness of individuals in each group along with the balancing of the number of each group in updating ($P_g$). As shown in Equations (1) and (2) we have developed a two-group strategy as discussed in this paper with subsequent sections presenting the experimental design and implementation.

The ABGS is designed as follows: initially, the 1st, 2nd, ... 10th individual are assigned randomly to the groups (each

TABLE 1
Parameter of Four Standard Test Functions

| *Function* | *Expression* | *Dimension* |
|---|---|---|
| Sphere | $f_1 = \sum_{i=1}^{d} x_i^2$ | 30 |
| Rosenbrock | $f_2 = \sum_{i=1}^{d} \left[100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2\right]$ | 30 |
| Griewank | $f_3 = \frac{1}{4000}\sum_{i=1}^{d} x_i^2 - \prod_{i=1}^{d} \cos\left(\frac{x_i}{i^{1/2}}\right) + 1$ | 30 |
| Rastrigin | $f_4 = \sum_{i=1}^{d} [x_i^2 - 10\cos(2\pi x_i) + 10]$ | 30 |

group having 1 member at this stage). The 11th, 12th . . . 20th individuals are then assigned randomly to the groups (each group having 2 members at this stage). This iterative process is repeated until all individuals are assigned to groups. This grouping strategy effectively balances the individual fitness quality for each group thus balancing the number in each group updating ($Pg$).

Validation uses CGS and ABGS to test four standard test functions (the parameters for the four benchmark functions is shown in Table 1). The four benchmark functions are used to verify the performance of the algorithm with the experimental results being an average value obtained from 20 independent runs of three strategies. The experiment results are shown in Fig. 2.

In Fig. 2, *abscisca* [in a system of coordinates, the distance from a point to the vertical or $y$-axis, measured parallel to the horizontal or $x$-axis; the $x$–coordinate] expresses the serial number for each group, *ordinates* express the average update rate each group update ($Pg$).

An analysis identifies that the number [of individuals] in the first subgroup updating ($Pg$) is much higher than other groups with CGS, however, with the introduction of ABGS, the balance for each group update ($Pg$) was effectively retained thus avoiding premature convergence to local optima while achieving improved in the performance of the algorithm.

*3) ATF*: In order to represent the feature subset, the SFLA should be converted into a binary SFLA. The conversion formula as discussed in [22] is shown as (4), (5), new ($P_w$) is transformed into a vector in the binary range [0, 1] by:

$$sig(Dis) = \frac{1}{1 + e^{-A*Dis}} \qquad \mathrm{A} = \frac{g}{G}(factor1 - factor2) + factor2 \tag{4}$$

$$newP_w = \begin{cases} 1 & if(sig(Dis) > R) \\ 0 & if(sig(Dis) \le R) \end{cases}. \tag{5}$$

The new position of an individual is updated as follows:

- The adaptive transition factor ($A$) is introduced into conversion formula where: factor1 is 0.95, factor2 is 1.05, ($g$) is the current number of iterations, ($G$) is the total number of iterations, in the early stage of the algorithm.
- This (1) enhances the transition uncertainty of the conversioning linear solution to a discrete solution, (2) strengthens the ability to traverse the solution space, and (3) in the later stages of the algorithm implementation enhances the transition to improve fine grained searching when converging on the optimal value.

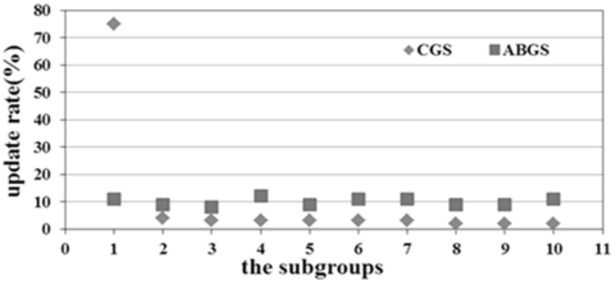


Fig. 2. The average update rate of each group updating $Pg$.

Thus the optimization performance of the algorithm is improved.

### 4.2. The Process of Feature Selection

In this paper, we apply a randomized ABGS grouping strategy with a cross operation and an adaptive transition factor to the basic SFLA. Two improvements to the SFLA are implemented in the proposed ISFLA. The details of the processes used to enable Feature Selection with ISFLA are as follows.

- *Step 1*: Randomly generate a population of F = m×n individuals. Where (***m***) represents the number of subgroups, (***n***) represents the number of individual in each subgroup, and each individual is converted to a binary number set by Equations (4) and (5).
- *Step 2*: Use the evaluation function [see Section 5] to calculate the value for each individual and identify the global best individual (***Pw***).
- *Step 3*: Using ABGS, sort (***F***) individuals into descending order and assign them to (***m***) subgroups where each subgroup contains (***n***) individuals.
- *Step 4*: Find the best (most fit) individual (***Pb***) and the worst (least fit) individual (***Pw***) in each subgroup. Up date (***Pw***) in each subgroup by Equations (2), (3). By cross operation the new (***Pw***) is converted to binary (***Pw***) by Equations (4), (5).
- *Step 5*: Calculate the evaluation function value for each individual and find the globally fittest individual (***Pg***).
- *Step 6*: Where the hybrid iteration number reaches (***n***), repeat the sorting and grouping of all individuals.
- *Step 7*: Test if the stopping criterion is met. Where the stopping criterion is met terminate the program run and output the optimum (***Pg***). Where the stopping criterion is not met, return to step 3.

The process of feature selection with ISFLA is presented in flow chart shown in Fig. 3 where ($S$) is equal to the number of algorithms executed in each experiment, its value is set in Table 3.

## 5 Evaluation Function

In our proposed method, classification accuracy and the number of selected features are the two indicators used to design the evaluation function as defined in [22]:

$$fitness = w_1 \times acc(Knn) + w_2 \times \left(1 - \frac{n}{N}\right). \tag{6}$$

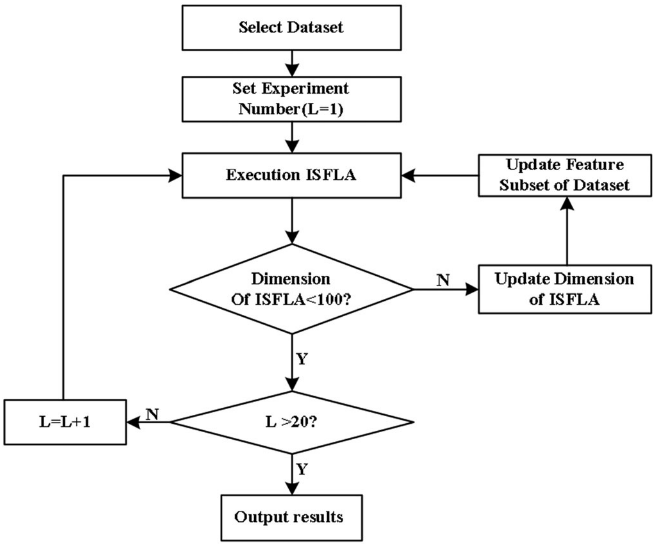

Select Dataset
Set Experiment Number(L=1)
Execution ISFLA
Update Feature Subset of Dataset
Dimension Of ISFLA<100?
N
Update Dimension of ISFLA
Y
L=L+1
N
L >20?
Y
Output results

Fig. 3. The feature selection flow chart.

As defined by Equation (6) the fitness functon has two predefined weights: ($w_1$) (the classification accuracy) and ($w_2$) (the selected feature). The accuracy [of the weight] can be adjusted to a high value if accuracy is the most important factor, in this paper, the values for ($w_1$) and ($w_2$) are [1] and [0.001] respectively. Given that an individual with a high fitness value has a high probability of effecting other individuals' positions in the next iteration, the weights ($w_1$) and ($w_2$) must be appropriately defined. (*acc*) is the classification accuracy where (*n*) is the number of select features and (*N*) is the total number of features.

For the fitness definition, (*acc*) [or hit rate] denotes the percentage of correctly classified examples as evaluated by Equation (7). The numbers of correctly and incorrectly classified examples are indicated by (*numc*) and (*numi*) respectively.

$$acc = \frac{num_c}{num_c + num_i} \times 100\%. \tag{7}$$

## 6 Results and Discussion

In this section we present an evaluation of the performance of our proposed approach. The evaluation uses a number of well known and recognized biomedical datasets [28]. The datasets include: *ColonTumor*, *DLBCL-Harvard*, and *Nervous-System* etc and provide data relating to gene expression, protein profiling, and genomic sequence for classification and disease diagnosis. All the datasets are high-dimensional and include less instance and irrelevant or weak correlation features, the dimensiona scope is from 2,000 to 12,600. These data sets are taken from the Kent Ridge Bio-medical Dataset (http://datam.i2r.a-tar.edu.sg/datasets/krbd/index.html), the format of datasets are showed in Table 2.

### 6.1 Parameter Setting

To evaluate of our proposed ISFLA algorithm we have conducted a comparative analysis using a SFLA, improved GA in [18] and the improved PSO in [22] methods selected for feature selection performance comparison. In our experiments, the consistent conditions and parameters are used in the comparative analysis where: the population size is 20, the number of iterations is 50, and the 1-nearest neighbor method is used to evaluate feature subsets. In ISFLA and SFLA the values for (m) and (n) are 4 and 5 respectively.

TABLE 2
The Format of Datasets

| *Dataset* | *Instances* | *Attributes* | *Classes* | *1-nn* |
|---|---|---|---|---|
| ColonTumor | 62 | 2000 | 2 | 75.80(0.25) |
| DLBCL-Harvard | 58 | 7129 | 2 | 46.55(0.53) |
| Nervous-System | 60 | 7129 | 2 | 56.67(0.43) |
| LungCancer-Harvard1 | 203 | 12600 | 5 | 89.66(0.05) |
| ALL-AML-Leukemia | 106 | 7130 | 2 | 89.47(0.12) |
| LungCancer-Ontario | 39 | 2880 | 2 | 56.41(0.44) |
| DLBCL-Stanford | 47 | 4026 | 2 | 76.60(0.25) |
| DLBCL-NIH | 160 | 7400 | 2 | 48.13(0.52) |
| LungCancer-Harvard2 | 181 | 12534 | 2 | 95.58(0.05) |

To demonstrate the generalization capability, the training and the test samples should be independent. In our experimentation we use 10-fold cross validation to estimate the classification rate on each dataset. The data are stratified into 10 folds. For the 10 folds, 9 folds constitute the training set with the remaining fold being used as the test set.

To avoid bias, all results are the average value for the algorithm independently executed 20 times; the objectives are to reduce the number of dataset feature subsets to less than 100 and improve the classification accuracy of the data-set(s). Nine typical high-dimensional biomedical data sets were selected (see Table 2), the column headed *1-nn* in Table 2 shows the classification accuracy of original data set and the data in brackets expresses the mean absolute error.

### 6.2. Results Analysis and Discussion

In Table 3 we list each dataset and the algorithms applied in the comparative analysis. Against each algorithm there are six attributes tabulated which are: (1) the average fitness (*Avg%*), (2) the highest fitness (*Max*), (3) the lowest fitness (*Min%*), (4) the standard deviation (**std**), (5) the average number of feature subsets (*AvgN*), and (6) the number of algorithm executions in each experiment (*S*).

It can be seen from Table 3 that the results produced by ISFLA achieved the best *Avg* among all the four algorithms for seven out of nine data sets. The *Avg* for ColonTumor, DLBCL-Harvard, Nervous-System, LungCancer-Ontario, DLBCL-Stanford, DLBCL-Stanford and DLBCL-NIH obtained by the ISFLA are 93.02, 73.33, 81.67, 75.06, 82.67, 55.63 and 98.89 percent, respectively. For the data sets: LungCancer-Harvard1 and ALL-AML-Leukemia, while the PSO obtained the best *Avg* at 91.90, and 99.01 percent, respectively, the *AvgN* for the two datasets with PSO are 44.20, 113.5, respectively which is much larger then is the case for the ISFLA.

In terms of the number of feature subsets and the *AvgN* metrics, the ISFLA obtained the smallest *AvgN* for all Datasets as compared to the SFLA, IGA and IPSO algorithms. We may also observe that The Std metric among all the four algorithms for seven out of nine data sets [as obtained by the ISFLA] is smaller than is the case for the other three algorithms evaluated. The best attributes values are highlighted and bold in Table 3.

Table 4 shows three average attribute values (avg(Avg), avg(Std), and avg(AvgN)) for nine datasets using the four

TABLE 3
The Result of Four Algorithm

| *Dataset* | *Algorithm* | *Avg(%)* | *Max(%)* | *Min(%)* | *Std* | *AvgN* | *S* |
|---|---|---|---|---|---|---|---|
| ColonTumor | ISFLA | **93.02** | 96.67 | 90.11 | 2.74 | **35.22** | 6 |
| | SFLA | 89.03 | 91.67 | 85.02 | **2.11** | 36.16 | 6 |
| | IGA | 86.67 | 88.33 | 83.33 | 2.36 | 38.24 | 6 |
| | IPSO | 87.67 | 91.67 | 85.01 | 3.65 | 49.40 | 6 |
| DLBCL-Harvard | ISFLA | **73.33** | 76.67 | 68.33 | **3.54** | **27.42** | 8 |
| | SFLA | 69.21 | 75.20 | 65.33 | 3.84 | 51.43 | 8 |
| | IGA | 64.33 | 70.06 | 60.00 | 5.21 | 27.62 | 8 |
| | IPSO | 71.11 | 76.67 | 63.33 | 5.34 | 51.24 | 8 |
| Nervous-System | ISFLA | **81.67** | 85.06 | 78.33 | 3.33 | **32.33** | 8 |
| | SFLA | 76.08 | 80.05 | 71.67 | **3.24** | 57.86 | 8 |
| | IGA | 71.67 | 81.67 | 61.67 | 7.16 | 30.25 | 8 |
| | IPSO | 72.67 | 78.33 | 63.33 | 6.07 | 45.03 | 8 |
| LungCancer-Harvard1 | ISFLA | 90.30 | 91.21 | 88.50 | **1.09** | **28.28** | 9 |
| | SFLA | 91.20 | 92.23 | 89.21 | 1.25 | 54.69 | 9 |
| | IGA | 85.90 | 87.50 | 84.09 | 1.29 | 31.81 | 9 |
| | IPSO | **91.90** | 94.14 | 90.04 | 1.51 | 44.20 | 9 |
| ALL-AML-Leukemia | ISFLA | 98.91 | 100.00 | 98.18 | **0.76** | **30.44** | 8 |
| | SFLA | 97.27 | 99.09 | 94.55 | 1.93 | 45.65 | 8 |
| | IGA | 95.09 | 97.27 | 92.73 | 1.65 | 30.63 | 8 |
| | IPSO | **99.01** | 100.00 | 98.18 | 1.04 | 113.5 | 8 |
| LungCancer-Ontario | ISFLA | **75.06** | 80.33 | 70.23 | **3.68** | **14.33** | 8 |
| | SFLA | 70.22 | 85.12 | 62.54 | 4.84 | 18.46 | 8 |
| | IGA | 65.50 | 75.21 | 57.51 | 4.18 | 10.22 | 8 |
| | IPSO | 70.00 | 77.50 | 57.50 | 4.89 | 56.25 | 8 |
| DLBCL-Stanford | ISFLA | **82.67** | 83.30 | 79.20 | 2.11 | **15.24** | 8 |
| | SFLA | 80.01 | 82.01 | 78.04 | 2.06 | 25.67 | 8 |
| | IGA | 78.80 | 84.02 | 72.02 | 4.82 | 18.43 | 8 |
| | IPSO | 78.10 | 80.02 | 74.31 | 3.16 | 49.50 | 8 |
| DLBCL-NIH | ISFLA | **55.63** | 56.88 | 52.50 | **2.10** | **29.25** | 8 |
| | SFLA | 54.06 | 58.13 | 50.63 | 3.13 | 30.75 | 8 |
| | IGA | 56.01 | 61.25 | 51.87 | 3.86 | 32.21 | 8 |
| | IPSO | 55.10 | 65.01 | 47.50 | 9.01 | 35.10 | 8 |
| LungCancer-Harvard2 | ISFLA | **98.89** | 99.44 | 97.77 | **0.78** | **51.50** | 8 |
| | SFLA | 98.01 | 98.79 | 96.67 | 1.06 | 75.25 | 8 |
| | IGA | 96.67 | 98.33 | 95.56 | 1.11 | 52.80 | 8 |
| | IPSO | 96.36 | 99.98 | 93.33 | 2.33 | 98.31 | 8 |

algorithms evaluated. In a comparative analysis comparing ISFLA with SFLA, IGA and IPSO we find that ISFLA demonstrates greater improvement in performance with better classification accuracy and stability while using fewer relevant feature subsets.

It can also be observed that due to the introduction of the CMWF, The ABGS and the ATF, the ISFLA explores the space of possible subsets to obtain the set of features that maximizes the predictive accuracy and minimizes irrelevant features in high-dimensional biomedical data.

The process of averaging the reducing value for feature subsets is shown in Figs. 4, 5, 6, 7, 8, 9, 10, 11, 12. In each figure, *abscissa* represents the number of feature subsets, and *ordinate* represents the average classification accuracy for each algorithm independently executed 20 times. Figs. 6, 8, 9, 12, present a performance comparision between the ISFLA and the SFLA, improved GA and improved PSO methods. Figs. 4, 5, 7, 10, 11, show that, while there is not a clear advantage in the early-middle stages, the ISFLA algorithm identified fewer feature subsets with a higher classification effect and better performance in the later stages.

TABLE 4
The Average Attributes Value of Nine Datasets Using Four Algorithms

| *attributes* | *ISFLA* | *SFLA* | *IGA* | *IPSO* |
|---|---|---|---|---|
| *Avg (Avg)* | 83.26 | 80.53 | 77.84 | 80.19 |
| *Avg (Std)* | 2.23 | 3.05 | 3.96 | 4.11 |
| *Avg (AvgN)* | 29.33 | 43.95 | 30.22 | 60.22 |

Considering Tables 3 and 4 with Figs. 4, 5, 6, 7, 8, 9, 10, 11, 12, we find that the three improvement strategies: (1) the ABGS, (2) the CMWF, and (3) the ATF (see Section 4.1) play an important role in the feature selection performance of ISFLA.

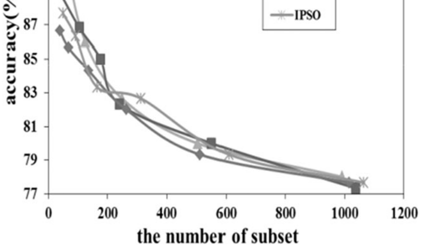


Fig. 4. The dimension reduction curve of ColonTumor.

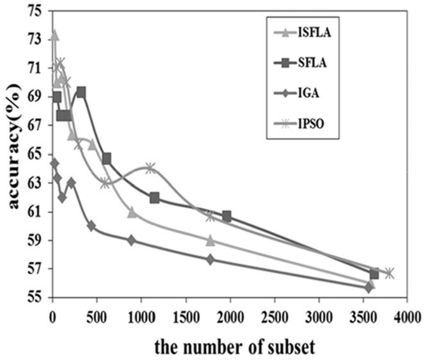


Fig. 5. The dimension reduction curve of DLBCL-Harvard.

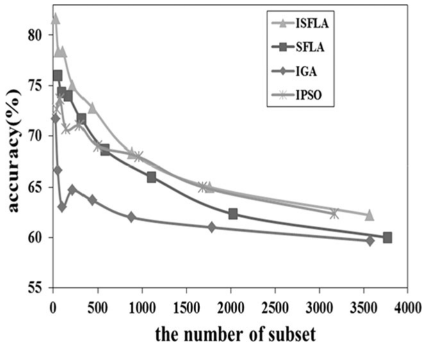


Fig. 6. The dimension reduction curve of Nervous-System.

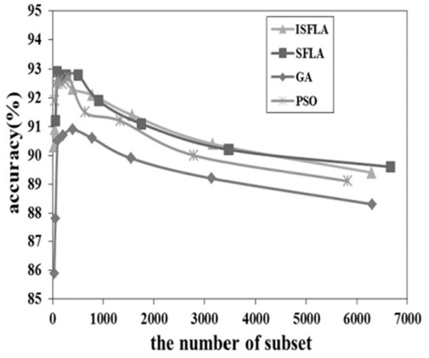


Fig. 7. The dimension reduction curve of LungCancer-Harvard1.

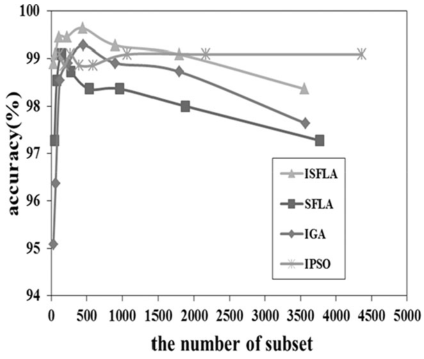


Fig. 8. The dimension reduction curve of ALL-AML-Leukemia.

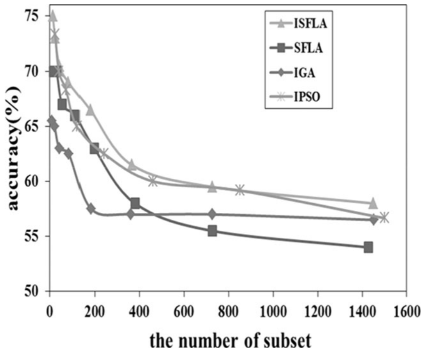


Fig. 9. The dimension reduction curve of LungCancer-Ontario.

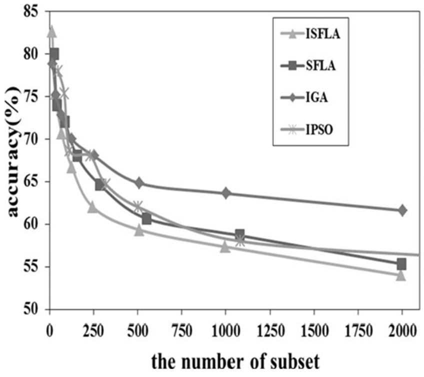


Fig. 10. The average dimension reduction curve of DLBCL-Stanford.

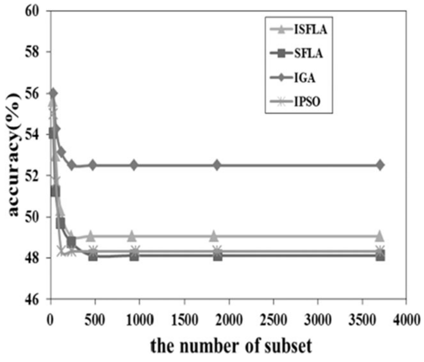


Fig. 11. The average dimension reduction curve of DLBCL-NIH.

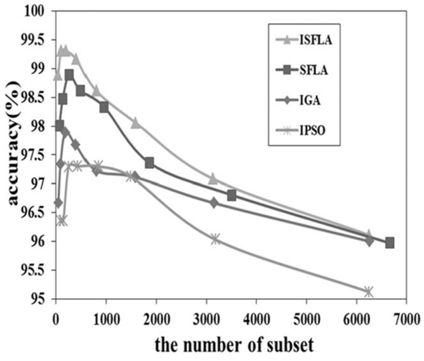


Fig. 12. The average dimension reduction curve of LungCancer-Harvard2.

It is worth noting that, the purpose of feature selection is to move unproductive features without reducing the predictive accuracy; otherwise, the performance might be degraded though that feature subset has small size.

For example, for Figs. 7, 8, 12, the average classification accuracy degraded gradually with the reduction in the number of features, therefore, we must balance the relationship between classification accuracy and the number of feature subsets in 'real-world' applications to make biological datasets play a more important role in disease diagnosis, and improve the effectiveness of disease diagnosis.

## 7 Conclusions

Feature subset selection is a fundamental technique in many application areas [12] and different evolutionary algorithms have been developed for different feature subset selection problems. In this paper, the SFLA algorithm is, for the first time, being used to solve a feature selection problem. By introducing the CMWF, The ABGS and the ATF, a new improved SFLA, termed ISFLA, is used to solve feature selection in high-dimensional biomedical data, and a 1-nearest neighbor (1-NN) serves as an evaluator of our proposed algorithms.

Experimental results show that our method effectively reduces the number of dataset features whilst simultaneously achieving a higher classification accuracy. The proposed method can serve as an ideal pre-processing tool to help optimize the feature selection process of high-dimensional biomedical data, better mine the function of biological databsets in fields of disease diagnosis, and improve the efficiency of disease diagnosis.

## Acknowledgments

Bin Hu is the corresponding author. Yongqiang Dai is the co-first author. This work is supported by the National Basic Research Program of China (973 Program) (no. 2014CB744600), the National Natural Science Foundation of China (nos. 61402211, 61063028 and 61210010), the Natural Science Foundation of Gansu Province (No. 1506RJZA007), and the Natural Science Foundation of the Higher Education Institutions of Gansu Province, China (2015A-008).

## References

[1] Y. Lu and J. Han, “Cancer classification using gene expression data,” *Inf. Syst.*, vol. 28, pp. 243–268, 2003.

[2] J. Misra, W. Schmitt, D. Hwang, L. Hsiao, S. Gullans, and G. Stephanopoulos, “Interactive exploration of microarray gene expression patterns in a reduced dimensional space,” *Genome Res.*, vol. 2, pp. 1112–1120, 2002.

[3] K. Lee, Z. Man, D. Wang, and Z. Cao, “Classification of bioinformatics dataset using finite impulse response extreme learning machine for cancer diagnosis,” *Neural Comput. Appl.*, vol. 22, pp. 457–468, 2013.

[4] M. Eusuff and K. E. Lansey, “Optimization of water distribution network design using the shuffled frog leaping algorithm,” *Water Res. Planning Manage.*, vol. 3, pp. 210–225, 2003.

[5] H. M. Hasanien, “Shuffled frog leaping algorithm for photovoltaic model identification,” *IEEE Trans. Sustain. Energy*, vol. 6, no. 2, pp. 509–515, Apr. 2015.

[6] M. Shahriari-kahkeshi and J. Askari, “Nonlinear continuous stirred tank reactor (CSTR) identification and control using recurrent neural network trained shuffled frog leaping algorithm,” in *Proc. 2nd Int. Conf. Control Instrum. Autom.*, 2011, pp. 485–489.

[7] A. Alghazi, S. Z. Selim, and A. Elazouni, “Performance of shuffled frog-leaping algorithm in finance-based scheduling,” *J. Comput. Civil Eng.*, vol. 26, pp. 396–408, 2012.

[8] Q. K. Pan, L. Wang, L. Gao, and J. Li, “An effective shuffled frog-leaping algorithm for lot-streaming flow shop scheduling problem,” *Int. J. Adv. Manuf. Technol.*, vol. 52, pp. 699–713, 2011.

[9] I. Perez, M. Gomez Gonzalez, and F. Jurado, “Estimation of induction motor parameters using shuffled frog-leaping algorithm,” *Elect. Eng.*,vol. 95, pp. 267–275, 2013.

[10] Xuehui Luo, Ye Yang, and Xia Li, “Solving TSP with shuffled frog-leaping algorithm,” in *Proc. 8th Int. Conf. Intell. Syst. Des. Appl.*, 2008, vol. 3, pp. 228–232.

[11] J. Ebrahimi, S. H. Hosseinian, and G. B. Gharehpetian, “Unit commitment problem solution using shuffled frog leaping algorithm,” *IEEE Appl. Mathematics Comput.*, vol. 218, pp. 9353–9371, 2012.

[12] M. G. Gonzalez, F. J. R. Rodriguez, and F. Jurado, “A binary SFLA for probabilistic three-phase load flow in unbalanced distribution systems with technical constraints,” *Electr. Power Energy Syst.*, vol. 48, pp. 48–57, 2013.

[13] T. H. Huynh and D. H. Nguyen, “Fuzzy controller design using a new shuffled frog leaping algorithm,” in *Proc. IEEE Int. Conf. Ind. Technol.*, 2009, pp. 1–6.

[14] J. R. Vergara and P. A. Estévez, “A review of feature selection methods based on mutual information,” *Neural Comput. Appl.*, vol. 24, pp. 175–186, 2014.

[15] H. Handels, T. Roß, J. Kreusch, H. H. Wolff, and S. J. Pöppl, “Feature selection for optimized skin tumor recognition using genetic algorithms,” *Artificial Intell. Med.*, vol. 16, pp. 283–297, 1999.

[16] Y. Chtioui, D. Bertrand, and D. Barba, “Feature selection by a genetic algorithm.application to seed discrimination by artificial vision,” *J. Sci. Food Agriculture*, vol. 76, pp. 76–86, 1998.

[17] Y. Zhang, D. Gong, Y. Hu, and W. Zhang, “Feature selection algorithm based on bare bones particle swarm optimization,” *Neurocomputing*, vol. 148, pp. 150–157, 2015.

[18] C. S. Yang, L. Y. Chuang, Y. J. Chen, and C. H. Yang, “Feature selection using memetic algorithms,” in *Proc. 3rd Int. Conf. Convergence Hybrid Inf. Technol.*, 2008, pp. 416–423.

[19] M. M. Kabir, M. Shahjahan, and K. Murase, “A new hybrid ant colony optimization algorithm for feature selection,” *Expert Syst. Appl.*, vol. 39, pp. 3747–3763, 2012.

[20] C. L. Huang and J. F. Dun, “A distributed PSO-SVM hybrid system with feature selection and parameter optimization,” *Appl. Soft Comput.*, vol. 8, 2008, pp. 1381–1391.

[21] X. Wang, J. Yang, X. Teng, and W. Xia, “Richard Jensen, Feature selection based on rough sets and particle swarm optimization,” *Pattern Recognition Lett.*, vol. 28, pp. 459–471, 2007.

[22] L. Y. Chuang, H. W. Chang, C. J. Tu, and C. H. Yang, “Improved binary PSO for feature selection using gene expression data,” *Comput. Biol. Chemistry*, vol. 32, pp. 29–38, 2008.

[23] J. Luo and M. Chen, “Improved Shuffled Frog Leaping Algorithm and its multi-phase model for multi-depot vehicle routing problem,” *Expert Syst. Appl.*, vol. 41, pp. 2535–2545. 2014.

[24] L. Wang, and C. Fang, “An effective shuffled frog-leaping algorithm for multi-mode resource-constrained project scheduling problem,” *Inf. Sci.*, vol. 181, pp. 4804–4822, 2011.

[25] Xia Li, Jianping Luo, Minrong Chen, Na Wang, “An improved shuffled frog-leaping algorithm with extremal optimisation for continuous optimisation,” *Inf. Sci.*, vol. 192, pp. 143–151, 2012.

[26] S. M. Bala and R. Meenakumari, “Optimum generation scheduling using an optimum generation scheduling using an improved adaptive shuffled frog leaping algorithm,” in *Proc. Int. Conf. Cogn. Comput. Inf. Process.*, 2015, pp. 1–6.

[27] B. Alatas, E. Akin, and A. B. Ozer, “Chaos embedded particle swarm optimization algorithms,” *Chaos Solitons Fractals*, vol. 40, pp. 1715–1734, 2009.

[28] High-dimentional biomedical data sets. [Online]. http://datam.i2r.a-star.edu.sg/datasets/krbd/index.html, [access: 2015. 08. 01].

**Bin Hu** received the PhD degree in computer science from the Institute of Computing Technology, Chinese Academy of Science, China. Since 2009, he has been the dean of the School of Information Science and Engineering, Lanzhou University, China. He has been also IET fellow, IEEE co-chair SMC TC on cognitive computing, guest professor of ETH Zurich, Switzerland, member at Large of ACM China, and vice chair of International Society for Social Neuroscience (China Committee). His research interests include pervasive computing, psycho-physiological computing, and data modeling. He has published about 200 papers in peer reviewed journals, conferences, and book chapters including *Science* (Suppl.), the *Journal of Alzheimer's Disease*, the *PLoS One*, the *PLoS Computational Biology*, the *IEEE Transactions*, the *IEEE Intelligent Systems*, AAAI, BIBM, EMBS, CIKM, ACM SIGIR, etc. He has been involved in some big national/international funds schemes, e.g., EU FP7, Chinese ”973” as PI. He has also served as associate editor in peer reviewed journals on cognitive science and pervasive computing.

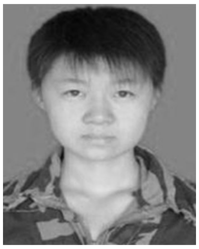
**Yongqiang Dai** received the BS degree in computer science and technology from Northwest Normal University, China, in 2004 and the MS degree in agricultural electrification and automation from Gansu Agriculture University, in 2011. He is currently working toward the PhD degree in computer application at Lanzhou University. Since 2014, he has been an associate professor in the School of Information Science and Technology, Gansu Agriculture University. His research interests include intersection between computer science and brain informatics, intelligent computing, and data mining.

**Yun Su** received the MS degree in computer software and theory from Sun Yat-sen University, China. She is currently working toward the PhD degree in computer application at Lanzhou University. She is a lecturer in the College of Computer Science and Engineering, Northwest Normal University, China. Her research interests include model checking, affective computing, and ontology-based knowledge base modeling of multimodal physiological signals.

**Philip Moore** received the BSc (Hons), MSc, and DEng degrees (PhD) from the Graduate School of Engineering, Fukuoka Institute of Technology, Japan. His research interests focus on intelligent contextaware systems in a range of domains including e-Healthcare and e-Learning systems. His work has been presented in international conferences and has been published in international computer science conference proceedings, journals, and books. He has served as a reviewer and a member of international program committees for international conferences.

**Xiaowei Zhang** received the BS degree in computer science and technology from Lanzhou University, China, in 2003 and the MS degree in computer application technology from Lanzhou University, in 2006. He is currently working toward the PhD degree in computer application at Lanzhou University. Since 2007, he has been a lecturer in the School of Information Science and Engineering, Lanzhou University. His research interests include intersection between computer science and brain informatics, including physiological computing, affective learning, and ubiquitous computing.

**Chengsheng Mao** received the BS degree in computer science and technology from the Huazhong University of Science and Technology, China, in 2008. He is currently working toward the PhD degree in computer application at Lanzhou University, China. His research interests include data mining, machine learning, affective computing, and bioinformatics.



**Jing Chen** is currently working toward the PhD degree in computer science and technology at Lanzhou University, China. Her research interest includes affective computing, multimodal fusion of physiological signals, and ontological modeling of physiological signals for emotion recognition.

**Lixin Xu** received the BS degree in information and computing science from Lanzhou University, Lanzhou, China, in 2013. He is currently working toward the MS degree in software engineering from Lanzhou University, China. His research interests include intersection between computer science and brain informatics, including physiological computing, affective learning, ubiquitous computing, and data mining. For example, the improvement of SVMs' performance, modeling of affective computing and prediction of psychiatric disorders.