Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 07-2022
DOI 10.17605/OSF.IO/AH3PB

BINARY SWALLOW SWARM OPTIMIZATION (BSSO) BASED FEATURE SELECTION AND DEEP AUTO-ENCODER BASED DATA CLUSTERING (DAEDC) FOR CLASSIFICATION USING PSYCHOLOGICAL SNP GENOMICS DATA

[1]Dr. R. KAVITHA, [2]Dr. R. RAJESWARI, [3]Dr. P. G. SIVAGAMINATHAN and [4]A. NITHYA

[1]Associate Professor, Department of Computer Science and Engineering, Dr. MGR Educational and Research Institute, Chennai
[2]Associate Professor, Department of Computer Science, Dr. MGR Educational and Research Institute, Chennai
[3]Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Educational Foundation,Vaddeswaram, Andhra Pradesh
[4]Assistant Professor, PSGR Krishnammal College for Institute, Chennai,Coimbatore

ABSTRACT:
Neurological disorders are clinically specified as infirmity that affects the mind,spinal cord. and the nerves at other parts the human body. The Structural, biochemical or electrical abnormalities in the mind, spinal cord or other nerves can result in quite a number of signs and symptoms. Regardless of its numerous perspectives,side effects, signs and effects, various studies have been made to distinguish the main cause of the problem. Genetic Association Studies have been a popular approach for assessing the association between common Single Nucleotide Polymorphisms (SNPs) and complex diseases. However, other genomic data involved in the Genetic Association Studies, for example, gene expressions, are usually neglected in these association studies. Recently, many Machine Learning algorithms have been utilized to identify the significant of SNP. The curseof dimensionality is the main challenge. On the otherhand, the number of samples is decidedly smaller than the number of SNPs. In addition, the number ofhealthy and patient samples can be unequal. These challenges make the feature selection and classification very difficult. Therefore, an efficient method is proposed to identify significant SNPs andclassify healthy and patient samples. Searching for the (sub) optimal subset of features is a Nondeterministic Polynomial Time (NP) hard problem. In this regard, firstly, the Mean Encoding, as an intelligent method, is utilized to convert the nominal SNP data to numeric. Then a Binary Swallow Swarm Optimization (BSSO) method is used for feature selection, which removes the irrelevant and redundant features. The binarization of the continuous swallow swarm meta-heuristics iscarried out using a special function. Finally, theproposed Deep Auto-Encoder Based DataClustering (DAEDC) algorithm is employed to classify so that it can construct its structure based on input data, automatically. To evaluate, apply the proposed approach to mental retardation SNPdataset, which obtained from the Gene Expression Omnibus (GEO) dataset. The proposed method has given higher results in terms of precision, recall, F-measure, and accuracy in mental retardation, and autism. The results indicate that it has succeeded with high efficiency, comparedwith other classifiers.

Keywords: Single Nucleotide Polymorphism (SNP), Feature selection, Complex diseases, BinarySwallow Swarm Optimization (BSSO), Deep learning, and Deep Auto-Encoder Based Data Clustering (DAEDC).

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 07-2022
DOI 10.17605/OSF.IO/AH3PB

1. INTRODUCTION

The availability of high-throughput genotyping technologies has greatly advanced biomedical research, enabling us to detect genetic variations that are associated with the risk of diseases with much finer resolution than before. With genome-wide genotyping of single nucleotide polymorphisms (SNPs) in the human genome, it is possible to evaluate disease-associated SNPs for helping unravel the genetic basis of complex genetic diseases [1]. SNPs are single nucleotide variations of DNA base pairs, and it has been well established in the Genome Wide Association Studies (GWAS) field that SNP profiles characterize a variety of diseases. In light of emerging research on GWAS, hundreds or thousands of objects (with disease or normal controls) are collected; each object is genotyped at up to millions of SNPs. This is a typical problem of the number of SNPs is typically thousands of times larger than the number of objects. The task is to identify genetic susceptibility of SNPs through assaying and analyzing SNPs at the genome wide scale [2]. In the SNP dataset, psychology deals by means of the cognitive, emotional, and behavioral aspects of the living organisms; thus, neurological disorders provide increase to psychological problems. A number of methods for analyzing of susceptibility of SNPs for neurological disorders in GWAS have been proposed in the literature, where each SNP is analyzed individually [3]. However, it is found that only a small portion of the SNPs have main effects on the complex disease traits, but most of the SNPs have shown little penetrance individually. On the other hand, many common diseases in humans have been shown to be caused by complex interactions among multiple SNPs. This is known as multilocus interactions [4]. For dealing with the later challenge, one way of testing the interactions is to exhaustive search the interactions between all SNPs.

However, some principal obstacles are challenging in the field of SNPs detection. The curse of dimensionality is the main challenge because the dimension of SNP data is very high (up to one million). In high dimensional data, typically, many features are irrelevant or redundant; these properties decrease the performance of the classifier and increase the computational cost. Additionally, the number of samples (healthy or patient) is decidedly smaller than the number of SNPs that means the SNP data are sparse. Also, the amount of healthy and patient samples can be unequal, which means the SNP data are unbalanced. Data with sparse and unbalance properties are other difficulties in most studies. Besides, we need to convert nominal data to numeric data in the SNP dataset, and which encoding method for this purpose must be used. Considering all these factors, the improvement of an efficient algorithm, involving feature selection and classification, is hard and complicated.

Thus, Feature Selection (FS) algorithms play an essential role because these algorithms can identify irrelevant and redundant features so that they can reduce dimensionality by removing these features. There are many FS algorithms that each algorithm can be suitable for each particular data. Therefore, discover which of them can be the best for each specific SNP data and will be able to select significant SNPs that cause to separate

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 07-2022
DOI 10.17605/OSF.IO/AH3PB

the healthy and patient samples with high accuracy. Feature selection techniques are designed to identify SNPs associated with complex traits. By selecting a reduced number of SNPs with significantly larger effects compared to other SNPs, researchers can focus on the most promising SNPs for use in genomic prediction. The reduced dimensionality provides for better generalisation due to a lower number of model parameters to be estimated from the data. Despite the importance of reducing dimensionality, only a few studies have used feature selection methods on GWAS data. The best subset of features contains the least number of dimensions that most contribute to prediction accuracy. FS is separate and different from model evaluation. It is therefore important to ensure that predictive models are evaluated on data that has not been used for estimating model parameters (training). This is commonly achieved by withholding a subset of data for testing once or repeatedly (e.g. in cross- validation).

Feature selection methods are divided into three groups: filters, wrappers, and embedded methods [6–7]. Filters are based on the generalized properties of training data and do not use any classifier construction algorithm in the process of feature selection. The advantages of this approach are relatively low computational complexity, sufficient generalization capability, and independence from the classifier. Its main disadvantage is that features are often selected independently [6]. Wrappers include the classifier construction procedure in the feature selection process and use the prognostic accuracy of the classifier to estimate the selected subset of features. The interaction with the classifier generally yields better results as compared to filters; however, it increases the computational complexity of the method and there is a risk of overfitting [6]. Embedded methods select features in the process of training and integrate the feature selection procedure into the classifier construction algorithm. Unfortunately, in some prediction studies there has been a tendency to select markers associated with an outcome using the complete dataset.

The feature selection problem can be modeled as a binary optimization problem [8], which is a Nondeterministic Polynomial time (NP) hard problem. Its optimal solution is guaranteed only by exhaustive search. Meta- heuristic methods allow one to find sub-optimal solutions of this problem without exploring the entire solution space. However, most meta-heuristics were designed for a continuous search space. A binary search space poses the problem of discontinuity and non- differentiability, which makes it difficult to use classical deterministic optimization methods [9].

Additionally, need a powerful model to classify SNP data into healthy and patient groups. The classification task is carried out based on the significant SNPs, which are selected by the FS algorithm. The performance of classification illustrates that the selected SNPs have a meaningful impact on complex diseases or not. In this study, focus on deep learning method. A deep learning algorithm is a specific subfield of the representation learning procedure, which detects multiple levels of representation [10,11,12]. High- level representation (or features) illustrates more aspects of the data [13,14]. In this paper, a

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 07-2022
DOI 10.17605/OSF.IO/AH3PB

novel method is proposed for feature selection based on Binary Swallow Swarm Optimization (BSSO). The binarization of the continuous swallow swarm metaheuristics is carried out using a special function. Finally, the proposed deep auto-encoder is employed to classify so that it can construct its structure based on input data, automatically. To evaluate, apply the proposed approach to mental retardation SNP dataset, which obtained from the Gene Expression Omnibus (GEO) dataset.

2. LIETRATURE REVIEW

Wu et al [15] developed an equal-width discretization scheme for informativeness to divide SNPs into multiple groups. In feature subspace selection, randomly select the same number of SNPs from each group and combine them to form a subspace to generate a decision tree. The advantage of this stratified sampling procedure can make sure each subspace contains enough useful SNPs, but can avoid a very high computational cost of exhaustive search of an optimal *mtry* , and maintain the randomness of a random forest. For Parkinson data also show some interesting genes identified by the method, which may be associated with neurological disorders for further biological investigations.

Nguyen et al [16] proposed to use a new Two-stage quality-based sampling method in Random Forests (ts-RF), for SNP subspace selection for GWAS. The method first applies p- value assessment to find a cut-off point that separates informative and irrelevant SNPs in two groups. The informative SNPs group is further divided into two sub-groups: highly informative and weak informative SNPs. When sampling the SNP subspace for building trees for the forest, only those SNPs from the two sub-groups are taken into account. The feature subspaces always contain highly informative SNPs when used to split a node at a tree. This approach enables one to generate more accurate trees with a lower prediction error, meanwhile possibly avoiding overfitting. It allows one to detect interactions of multiple SNPs with the diseases, and to reduce the dimensionality and the amount of Genome-wide association data needed for learning the RF model.

Kotlarz et al [17] built a tool, which uses array-based genotype information to classify next- generation sequencing–based SNPs into the correct and the incorrect calls. The deep learning algorithms were implemented via Keras. The results showed that for a rare event classification problem, like incorrect SNP detection in NGS data, the most parsimonious naïve model and a model with the weighting of SNP classes provided the best results for the classification of the validation dataset. Also explored the continuous [0, 1] space of the distribution of SNP class probability estimated by the deep learning network in order to determine the best cutoff for SNP binary classification.

Boutorh and Guessoum [18] proposed a new hybrid intelligent technique based on Association Rule Mining (ARM) and Neural Networks (NN) which uses Evolutionary

Algorithms (EA) is proposed to deal with the dimensionality problem. On the one hand, ARM optimized by Grammatical Evolution (GE) is used to select the most informative

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 07-2022
DOI 10.17605/OSF.IO/AH3PB

features and to reduce the dimensionality by parallel extraction of associations between SNPs in two separate datasets of case and control samples. On the other hand, and to complement the previous task, a NN is used for efficient classification. The Genetic Algorithm (GA) is used for setting up the parameters of the two combined techniques. The proposed GA-NN- GEARM approach has been applied on four different SNP datasets obtained from the NCBI Gene Expression Omnibus (GEO) website.

Aluzbi et al [19] proposed an accurate hybrid feature selection method for detecting the most informative SNPs and selecting an optimal SNP subset. The proposed method is based on the fusion of a filter and a wrapper method, i.e., the Conditional Mutual Information Maximization (CMIM) method and the Support Vector Machine- Recursive Feature Elimination (SVM-RFE) respectively. In general, from these results we conclude that SNPs of the whole genome can be efficiently employed to distinguish affected individuals with complex diseases from the healthy ones.

Pirmoradi et al [20] proposed firstly, the Mean Encoding, as an intelligent method, is utilized to convert the nominal SNP data to numeric. Then a two-step filter method is used for feature selection, which removes the irrelevant and redundant features. Finally, the proposed deep auto-encoder is employed to classify so that it can construct its structure based on input data, automatically. To evaluate, we apply the proposed approach to five different SNP datasets, including thyroid cancer, mental retardation, breast cancer, colorectal cancer, and autism, which obtained from the Gene Expression Omnibus (GEO) dataset.

Sun et al [21] proposed a parameterized graph-theoretic generalization model to SNP data as a similarity network and searched for representative SNP variables. In particular, each SNP was represented as a vertex in the graph, (dis)similarity measures such as correlation coefficients or pairwise linkage disequilibrium was estimated to describe the relationship between each pair of SNPs; a pair of vertices are adjacent, i.e. joined by an edge, if the pairwise similarity measure exceeds a user-specified threshold. A minimum k-dominating set in the SNP graph was then made as the smallest subset such that every SNP that is excluded from the subset has at least k neighbors in the selected ones. The strength of k- dominating set selection in identifying independent variables, and in culling representative variables that are highly correlated with others, was demonstrated by a simulated dataset.

3. PROPOSED METHODOLOGY

The proposed process involves three stages: (A) A pre-processing stage, which consists of encoding nominal SNP data with neurological disorder for psychological analysis, and removing or replacing missing values in SNP data. (B) A feature selection stage; in this stage, significant SNPs are selected by a suitable FS algorithm. (C) A classification stage, the self-organizing auto- encoders are utilized to classify SNP data in this stage.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 07-2022
DOI 10.17605/OSF.IO/AH3PB

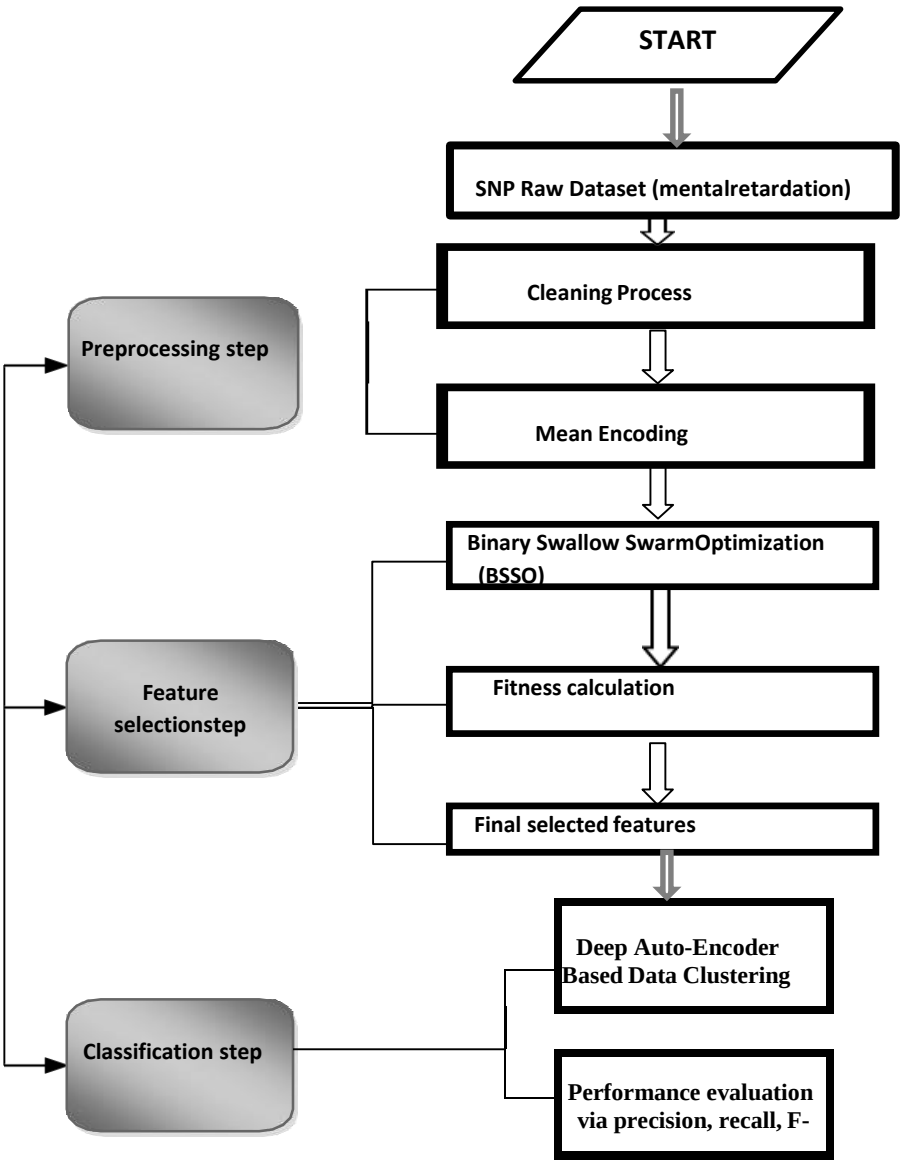

FIGURE 1. THE OVERALL FLOW OF THE PROPOSEDSYSTEM

Also, selected SNPs are evaluated according tosome classification metrics such as Precision,

Also, selected SNPs are evaluated according tosome classification metrics such as Precision, Recall, F-measure, and accuracy. The wholeprocess is shown in figure 1.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 07-2022
DOI 10.17605/OSF.IO/AH3PB

3.1 SNP dataset

All of the SNP data studied in this experiment are taken from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) website. The proposed method is carried out on complex disease from GSE13117 which consist of 250,000 SNP with total number of samples as 360 in which 1200belong to case and 240 belongs to control. The dataset consist of a different number of samples (case and control), a different number of features (SNPs), and various genetic patterns. In addition, these datasets are available in GEO of the National Center for Biotechnology Information, which is one of the most reliable biology data centers in the world.

3.2 Pre-processing

The pre-processing stage involves two steps. In the first step, redundant SNPs (or features) are detected and removed then in the second step, No Call or missing values are replaced by suitable amounts.

3.2.1 Redundant features

The SNPs that have the same values for all case and control samples are considered as redundant features since these features or SNPs cannot separate the two groups (case and control). For example, if BB value is registered for a given SNP in all case and control, it will not be helpful for feature selection and classification.

3.2.2 Missing values

Each SNP that involves more than 10% of No Call values are discarded; otherwise, the No Call value is replaced by estimated value, which in this study is the mode of feature (most common value for a given feature in all samples) in some SNP datasets or is considered as a new feature type in some other SNP datasets.

3.2.3 Encoding method

There are different types of data in data analysis. Generally, data can be assumed as numerical data and nominal data. Most Machine Learning algorithms utilize the mathematical calculations for feature selection and classification, so nominal data should be converted to a numerical amount that can be useful. In this study, all of the SNP data are nominal data (like BB, Bb, and bb); therefore, need a suitable encoding method to transform the SNP data to the numerical amount. There are various encoding methods [22], and each technique can be suitable for specific data. The binary and one hot encoding are used in the majority of studies [23]. However, these encoding methods increase the number of features based on the number of attribute categories. Also, encoding methods such as one hot, binary, integer representation, etc., which are used in recent studies, do not describe any information about the difference or similarity of SNP genotypes. In this study, the mean encoding is utilized for encoding the SNP data. The mean encoding method is an intelligent method since it can consider the target label in

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 07-2022
DOI 10.17605/OSF.IO/AH3PB

the encoding process, whereas other encoding methods have no correlation with the target. Also, mean encoding could prove to be a much simpler method in case of a large number of features. Mean encoding calculates the numerical amount for the unique nominal feature according to equation (1).

mean encoding of feature i=Number of true targets under the feature i/ Total number of targets under the feature i (1)

Therefore, applying the intelligent encoding method such as mean encoding, this applies useful information to label categories. In mean encoding, the algorithm assigns the quantity to each category (BB, Bb, and bb) based on the separability potential of each category in each specific data. In the raised example, the algorithm assigns the same quantity for BB and Bb that is equal to 0.6 (3 (case)/5 (instances)) in the i^{th} feature, which illustrates to be BB and Bb is not significant in the case group by assigning the same quantities. Therefore, the i^{th} feature can be irrelevant, and this property of the mean encoding can improve the performance of the feature selection step and classification step. Otherwise, mean encoding assigns the different quantities for BB and Bb, which illustrates to be BB or Bb in i^{th} feature play a significant role in complex disease as shown.

3.3 Wrapper-Based Feature Selection Algorithm BY Binary Swallow Swarm Optimization (BSSO)

Feature Selection (FS) plays an essential role in machine learning and pattern recognition; FS can improve the accuracy of a classifier by removing irrelevant and redundant features. The FS algorithm tries to select a subset of features from input data that can efficiently explain the input data and despite decreasing the effects of noise (irrelevant and redundant features) still can provide good classification and prediction results.

3.2.4 Swallow Swarm Optimization (SSO)

Swallow swarm optimization is a population-based metaheuristic based on the algorithm and it is used for feature selection [24]. At the beginning of each iteration, the population is sorted based on the value of the objective function. Then, the following roles are assigned:

1. Head leader is a particle with the best value of the objective function;
2. Local leaders are l particles that follow the head leader in accordance with the value of the objective function;
3. Aimless particles are k particles with the worst value of the objective function;
4. Explorers are all other particles.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 07-2022
DOI 10.17605/OSF.IO/AH3PB

On the current iteration, head leaders do not move, acting as beacons for explorer particles, which, in turn, explore the search space between the nearest local leader and the head leader. Explorer particles change their positions by the following equation (2-5),

$$\theta_e(t+1) = \theta_e(t) + V(t+1) \quad (2)$$

$$V(t+1) = V_{HL}(t+1) + V_{LL}(t+1) \quad (3)$$

$$V_{HL}(t+1) = V_{HL}(t) + rand(0,1)(\theta^{best}(t) - \theta_e(t)) + rand(0,1)(\theta_{HL}(t) - \theta_e(t)) \quad (4)$$

$$V_{LL}(t+1) = V_{LL}(t) + rand(0,1)(\theta^{best}(t) - \theta_e(t)) + rand(0,1)(\theta_{LL}(t) - \theta_e(t)) \quad (5)$$

where θ_e is the position of the explorer, θ_{HL} is the position of the head leader, θ_{LL} is the position of the local leader nearest to the explorer, θ^{best} is the best position, V is the velocity vector of the particle, V_{HL} is the velocity vector of the particle moving to the head leader, and V_{LL} is the velocity vector of the particle moving to the nearest local leader. In contrast to [24] decide not to select the parameters α_{HL}, β, α_{LL} and β_{LL} which are used to compute the velocity vectors with respect to the head and local leaders. Our experiments showed that the corresponding procedure increases the runtime of the algorithm without any significant improvement of the result. In this work, all these parameters are set to 1. The equation for changing the positions of aimless particles is also modified because the original equation can cause particles to gather at the boundaries of the search space or even go beyond it. Equation reduces the probability of this behavior and also allows explorer particles to slightly affect the behavior of aimless particles. To change the position of aimless particles, the following equations (6-7),

$$\theta_o(t+1) = rand(0.5,2) \cdot VSS \quad (6)$$

$$VSS = \frac{\sum_{j=1}^{N-k} \theta_{e(t)}^{j}}{N-k} \quad (7)$$

where θ_o is the position of aimless particle, θ^j is the position of the j-th particle, N is the total number of particles in the population, and k is the number of aimless particles. Once the termination condition is met, the algorithm returns the position of the head leader as a new solution.

3.2.5 Binary Swallow Swarm Optimization Algorithm

This section describes the proposed feature selection method based on the Binary Swallow Swarm Optimization (BSSO) algorithm. BSSO adapts the original algorithm to

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 07-2022
DOI 10.17605/OSF.IO/AH3PB

solve feature selection problems. This algorithm uses operators that can only be used to solve this particular problem. Here, the position is a Hamiltonian cycle, and the velocity is defined as a set of permutation between two cities [24] . In contrast to [25] in this work, use a special function merge to update the position of particles. As its input, this function receives two vectors $X\ and\ Y$, as well as the number , which determines the influence of $X\ and\ Y$, ranging from 0 to 1. The function merge yields the vector Z each element of which is found as follows: if the values $X_i\ and\ Y_i$ coincide, then has the same value; otherwise, takes the value X_i with the probability p or takes the value Y_i with the probability $(1 - p)$.

$$merge(X, Y, p)_i = \begin{cases} X_i, if\ rand(0,1) < p \\ Y_i\ otherwise \end{cases}, \quad i = 1,2, \ldots, D. \tag{8}$$

This is due to the fact that all solutions have different classification accuracies. If a feature is included in a solution that has higher classification accuracy, then this feature is more likely to be relevant. Conversely, if a feature is not included in a solution with higher classification accuracy, then this feature is more likely to be noise. Thus, this approach allows a new solution to include more features that are potentially relevant and to exclude more features that are potentially noise. On the other hand, by varying p, it is possible to control the effect of randomness when computing a new solution, so that the algorithm does not converge too quickly and is not stuck at local optima. This algorithm represents solutions as vectors S that encode features. At the first step of the algorithm, a population (set) of vectors S is generated (randomly or in some other way). The number of vectors in the population is a preset integer, which is also referred to as the population size. For each vector, a measure of classification accuracy is computed. On each iteration, all vectors are sorted in descending order of accuracy. The first element becomes the head leader. The next
i solutions are local leaders, n worst vectors are aimless particles, and all other vectors are explorer particles. The integer variables i and n are also specified beforehand. Explorer particles change their positions based on the positions of the leaders, while aimless particles do so randomly. Below are the corresponding equations (8-11) for explorers,

$$S_e(t + 1) = merge(V(t + 1), S_e(t), p_{ve}) \tag{8}$$

$$rge(V_{HL}(t + 1), V_{LL}(t + 1), p_{vhi}), \tag{9}$$

$$V_{HL}(t + 1) = merge(merge(S_{HL}(t), S_e(t), p_{he}), rand\{0,1\} \tag{10}$$

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 07-2022
DOI 10.17605/OSF.IO/AH3PB

$$V_{LL}(T+1) = merge(merge(S_{LL}(t), S_e(t), rand\{0,1\}^D, p \quad (11)$$

where $\boldsymbol{S_{HL}}$ is the position of the head leader,
$\boldsymbol{S_{LL}}$ is the position of the local leader, S_e is the position of the explorer, **VHL** is the velocity vector with respect to the head leader, **VLL** is the velocity vector with respect to the nearest local leader, V is the common velocity vector, p_{ve} is the effect of the velocity vector on the position of the explorer, p_{vhl} is the effect of the velocity vector with respect to the head leader on the velocity vector with respect to the local leader, p_{he} is the effect of the head leader's position on the position of the explorer, p_{her} is the combined effect of the head leader and explorer on a random vector, p_{le} is the effect of the local leader's position on the position of the explorer, and p_{ler} is the combined effect of the local leader and explorer on a random vector. A pseudo code of the BSSO is shown in Algorithm 1.

ALGORITHM 1. BINARY SWALLOW SWARM OPTIMIZATION (BSSO) BASED FEATURE SELECTION
INPUT: train data
OUTPUT: S_{HL} – position of the head leader
PARAMETERS: iterations – maximum number of cycles, N – population size, D – dimension, p_{vs} , p_{vhi}, p_{he}, p_{her}, p_{le}, p_{ler}, l– local leaders, k – aimless particles
1. Begin
2. for $i \leftarrow to\ N$ do
3. Initialize each solution S^i in the population: $S^i \leftarrow rand\ \{0,1\}^D$;
4. Select j^{th} (j = 1,2, … , D) feature for subset Sub where $sub\ where S^i$ = 1;

5. Build reduced dataset (subtra) based on Sub;

6. Evaluate fitness value ($fitness_i$) of Sub: $fitness_i \leftarrow errorrate$;
7. Evaluate fitness value ($fitness_i$) of Sub: $fitness_i \leftarrow errorrate$;
8. end
9. SHL ←Sk, where k =min ($fitness$) i=1,2…,N
10. for iter←1 to iterations do
11. for i ← to N do
12. Find nearest local leader SLL among population {$S1$, $S2$, … . , SN);
13. Evolve a new solution Se = {$f1$, $f2$, … , fD}
14. Select j-th feature (j = 1,2, . . . ,D) for subset Sub, where Sej = 1;
15. Build reduced dataset ($subtra$) based on sub;
16. Evaluate fitness value ($fitnessi$) of Sub: $fitnessi \leftarrow errorate$:
17. if $fitnessSe < fitnessi\ then$
18. $Si \leftarrow Se$;
19. $fitnessi \leftarrow fitnessSe$;
20. end
21. end

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 07-2022
DOI 10.17605/OSF.IO/AH3PB

22. SHL ←Sk, where k =min $(fitness)$; i=1,2…,N
23. end
24. end

3.3 Deep uto-Encoder Based Data Clustering (DAEDC)

Deep Auto-Encoder Based Data Clustering (DAEDC) algorithm is proposed in the data layer of a dataset with selected features is firstly mapped to the code layer, which is then used to reconstruct the data layer to classify the SNP data. The objective is minimizing the reconstruction error as well as the distance between data points and corresponding clusters in the code layer. This process is implemented via a four-layer auto-encoder network, in which a non-linear mapping is resolved to enhance data representation in the data layer. For clarity, here firstly introduce the auto-encoder network, and then explain how to use it for clustering.

3.3.1 Auto-encoders

Without loss of generality, take a one- layer auto-encoder network as an example. It consists of an encoder and a decoder. The encoder maps an input x_i to its hidden representation h_i.

The mapping function is usually non-linear and the following is a common form,

$$h_i = f(x_i) = \frac{1}{1+\exp(-(W_1 x_i + b_1))} \quad (12)$$

Where W_1 is the encoding weight, b_1 is the corresponding bias vector. The decoder seeks to
Reconstruct the input x_i from its hidden representation h_i. The transformation function has a similar formulation,

$$x_i' = h(h_i) = \frac{1}{1+\exp(-(W_2 x_i + b_2))} \quad (13)$$

where W_1, b_2 are the decoding weight and the decoding bias vector respectively. The auto- encoder model aims to learn a useful hidden representation by minimizing the reconstruction error. Thus, given N training samples, the parameters W_1 , W_2 , b_1 and b_2 can be resolved by the following

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 07-2022
DOI 10.17605/OSF.IO/AH3PB

optimization problem

$$\min \frac{1}{N} \sum_{i=1}^{N} \| x_i - x_i' \|^2 \tag{14}$$

Generally, an auto-encoder network is constructed by stacking multiple one layer auto-encoders. That is, the hidden representation of the previous one-layer auto-encoder is fed as the input of the next one [25].

3.3.2 Clustering Based on Auto-encoder

Auto-encoder is a powerful model to train a mapping function, which ensures the minimum reconstruction error from the code layer to the data layer. Usually, the code layer has less dimensionality than the data layer. Therefore, auto- encoder can learn an effective representation in a low dimensional space, and it can be considered as a non-linear mapping model. However, auto- encoder contributes little to clustering because it does not pursue that similar input data obtain the same representations in the code layer, which is the nature of clustering. To solve this problem, new objective function is embedded it into the auto-encoder model [26],

$$\min_{W,b} \frac{1}{N} \sum_{i=1}^{N} \| x_i - x_i' \|^2 - \lambda \sum_{i=1}^{N} \| f^t(x_i) - c^* \|_i^2 \tag{15}$$

$$c_i^* = \arg\min_{c_j^{t-1}} \| f^t(x_i) - c_j^{t-1} \|^2 \tag{16}$$

where N is the number of samples in the dataset; $f^t(\cdot)$ is the non-linear mapping function at the t^{th} iteration; c_j^{t-1} is the j^{th} cluster center computed at the $(t-1)^{th}$ iteration1; and c^* is the closest cluster center of the i^{th} sample in the code layer. This objective ensures that the data representations in the code layer are close to their corresponding cluster centers, and meanwhile the reconstruction error is still under control, which is important to obtain stable non-linear mapping. Two components need to be optimized: the mapping function $(\cdot)$ and the cluster centers c. To solve this problem, an alternate optimization method is proposed, which firstly optimizes $(\cdot)$ while keeps c fixed, and then updates the cluster center [26]

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 07-2022
DOI 10.17605/OSF.IO/AH3PB

4. RESULTS AND DISCUSSION

Briefly the following steps were done for all SNP data and then the simulation result were reported in this section. The existing classifiers and proposed system are implemented via the MATLAB simulation.

1- The SNP data was preprocessed so that features in which the number of missing values more than the determined threshold, user-defined parameters such as 10%, was removed and also the other missing values were replaced by suitable values.

$$c_j^t = \frac{\sum_{x_i \in C_j^{t-1}} f^t(x_i)}{|C_j^{t-1}|} \quad (17)$$

where C^{t-1}_j is the set of samples belonging to the j^{th} cluster at the $(t-1)^{th}$ iteration and $|C_j|$ is

the number of samples in this cluster. The sample assignment computed in the last iteration is used to update the cluster centers of the current iteration. Note that sample assignment at the first iteration C^0 is initialized randomly. Algorithm 2 represents the overall procedure of proposed Deep Auto-Encoder Based Data Clustering (DAEDC) algorithm.

ALGORITHM 2. DEEP AUTO-ENCODER BASED DATA CLUSTERING (DAEDC) ALGORITHM

INPUT : Dataset X, the number of clusters K, hyper-parameter λ, the maximum number of iterations T

OUTPUT : Final assignment C

Initialize sample assignment C^0 randomly

Set t to 1

Repeat

Update the mapping network by minimizing the equation (15) with stochastic gradient descent for one epoch

Update cluster center c^t via equation (16)

Partition X into K clusters and update the sample assignment C^t via equation (17)
t=t+1

Until t>T

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 07-2022
DOI 10.17605/OSF.IO/AH3PB

4. RESULTS AND DISCUSSION

Briefly the following steps were done for all SNP data and then the simulation result were reported in this section. The existing classifiers and proposed system are implemented via the MATLAB simulation.

1. The SNP data was preprocessed so that features in which the number of missing values more than the determined threshold, user-defined parameters such as 10%, was removed and also the other missing values were replaced by suitable values.
2. The SNP data were divided into three parts: namely training, validation, and test, involving 70%, 10%, and 20% of data, respectively.
3. The preprocessed SNP data were converted to numeric data using the Mean Encoding method.
4. FS algorithm is applied to SNP data (training data), in the first step, the FS algorithm calculates the relevance between each feature and target, and then 1000 top features were selected from dataset. Eventually, 100 top features (or SNPs) were chosen.
5. Classifier is applied to evaluate selected features. In this step, the proposed self-organizing deep auto-encoder was used to classify SNP data based on selected SNPs. The proposed DAEDC classifier can determine its structure automatically so that we did not require doing a random search or grid search to construct its structure. So, it has low time spending and low computational cost than the traditional method.

4.1 DATASET

The proposed method is carried out on complex disease from GSE13117 which consist of 250,000 SNP with total number of samples as 360 in which 1200belong to case and 240 belongs to control. Each centre tested DNA from 40 patients with unexplained mental retardation together with their parents. In addition, 38 DNA samples containing known submicroscopic copy number variations (CNVs) were run for validation purposes Keywords: genomic hybridisation Overall design: We performed a validation experiement where genomic DNA of 38 patients with mental retardation was hybridized onto Affymetrix' GeneChip 250K SNP (Nsp) arrays, and identified genome-wide CNVs. The dataset consist of a different number of samples (case and control), a different number of features (SNPs), and various genetic patterns. Precision, Recall, F-measure, and Accuracy metrics has been used to experiment the results of the classifiers.

4.2 EVALUATION MEASURES

The result of classification on SNP data (test data) indicates how much the proposed method can distinguish healthy and patient groups based on the selected SNPs. The prediction power of the self- organizing deep auto-encoder was evaluated using four measures such as accuracy, F- measure, Precision, and Recall which illustrates the ability of proposed method to predict cases, as shown in equations (18-21) respectively.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 07-2022
DOI 10.17605/OSF.IO/AH3PB

$$\text{Accuracy} = (TP + TN / TP + FN + FP + TN) * 100 \quad (18)$$

$$\text{easure} = 2(\text{Precision} \times \text{Recall} / \text{Precision} + \text{Recall}) \quad (19)$$

$$\text{Recall} = TP / TP + FN \quad (20)$$

$$\text{Precision} = TP / TP + FP \quad (21)$$

TP, FN, FP, and TN are the number of a True Positive, False Negative, False Positive, and False Negative respectively. The results of the proposed method (DAEDC) were compared with the results of other known Feature Selection (FS) methods such as the Minimum Redundancy Maximum Relevance (mRMR) algorithm [27], Conditional Mutual Information Maximin (CMIM) [28], and Relevance and Redundancy Analysis (RRA) [20] that 100 top SNPs were selected according to these methods. Also, popular classifiers such as Support Vector Machine (SVM), Support Vector Machine- Recursive Feature Elimination (SVM-RFE), and K-Nearest Neighbors (KNN) was utilized to evaluate the selected SNPs, then the accuracy and F-measure of above combination methods for SNP dataset.

TABLE 1. RESUTLS COMPARISON OF FSFOR CLASSIFIERS ON MENTAL RETARDATION DATASET

Metrics	Feature selection	KNN	SVM	SVM - RFE	DAEDC
Precision (%)	mRMR	81.12	83.25	86.21	88.17
	CMIM	83.27	85.18	90.17	91.21
	RRA	85.63	87.18	89.41	92.36
	BSSO	87.93	89.25	91.81	93.81
Re ca	mRM	85.1	87.5	89.1	92.51
	R	4	2	2	

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 07-2022
DOI 10.17605/OSF.IO/AH3PB

	CMIM	**87.63**	**90.47**	**91.36**	93.21
	RRA	**88.36**	**91.33**	**92.84**	94.18
	BSSO	**90.82**	**91.93**	**93.52**	94.43
F-measure (%)	**mRM R**	**84.62**	**86.47**	**89.34**	91.61
	CMIM	**85.92**	**87.92**	**90.37**	92.62
	RRA	**86.94**	**89.1**	**90.91**	93.21
	BSSO	**87.97**	**90.10**	**91.73**	93.63
Accuracy (%)	**mRM R**	**82.18**	**86.21**	**88.81**	89.75
	CMIM	**84.21**	**87.7**	**89.63**	90.15
	RRA	**86.71**	**88.21**	**90.14**	91.71
	BSSO	**88.18**	**90.17**	**91.36**	93.48

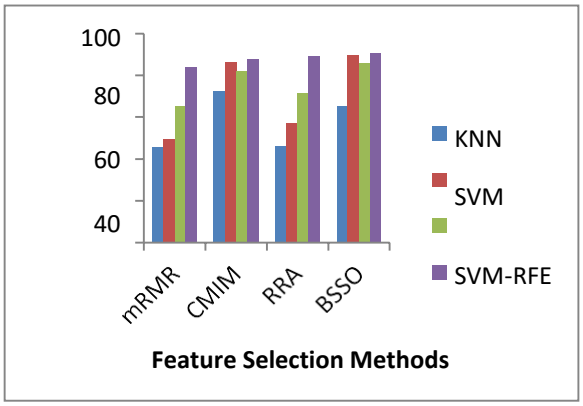

FIGURE 2. PRECISION RESULTS OF FS METHODS VS. CLASSIFIERS

Figure 2 proposed DAEDC classifier gives higher precision results for BSSO algorithm is 93.81%, similarly it also gives highest precision results for other FS methods such as mRMR, CMIM, and RRA methods like 88.17%,91.21%, and 92.36% respectively. The methods like KNN, SVM, SVM-RFE via BSSO algorithm also gives highest precision results of 87.93%, 89.25%, and 91.81% respectively(Refer Table 1).

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 07-2022
DOI 10.17605/OSF.IO/AH3PB

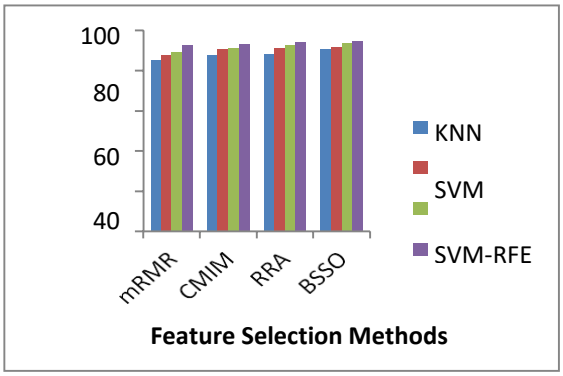

FIGURE 3. RECALL RESULTS OF FSMETHODS VS. CLASSIFIERS

Figure 3 proposed DAEDC classifier gives higher recall results for BSSO algorithm is 94.43%, similarly it also gives highest recall results for other FS methods such as mRMR, CMIM, and RRA methods like 92.51%,93.21%, and 94.18% respectively. The methods like KNN, SVM, SVM- RFE via BSSO algorithm also gives highest recall results of 90.82%,91.93%, and 93.52% respectively (Refer Table 1).

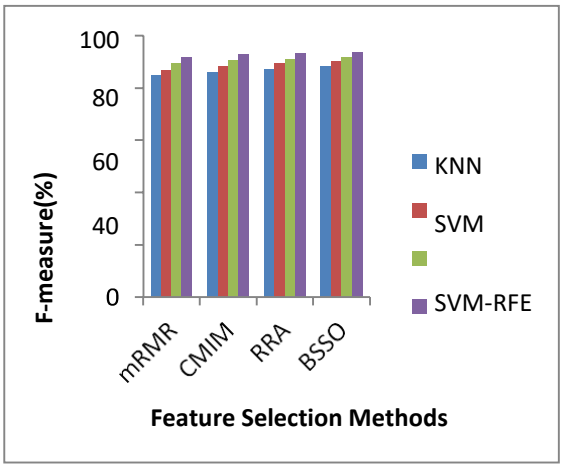

FIGURE 4. F-MEASURE RESULTS OF FSMETHODS VS. CLASSIFIERS

Figure 4 proposed DAEDC classifier gives higher f-measure results for BSSO algorithm is 93.63%, similarly it also gives highest f-measure results for other FS methods such as mRMR, CMIM, and RRA methods like 91.61%,92.62%, and 93.21% respectively. The methods like KNN, SVM, SVM-RFE via BSSO algorithm also gives highest f-measure results of 87.97%,90.10%, and 91.73% respectively (Refer Table 1).

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 07-2022
DOI 10.17605/OSF.IO/AH3PB

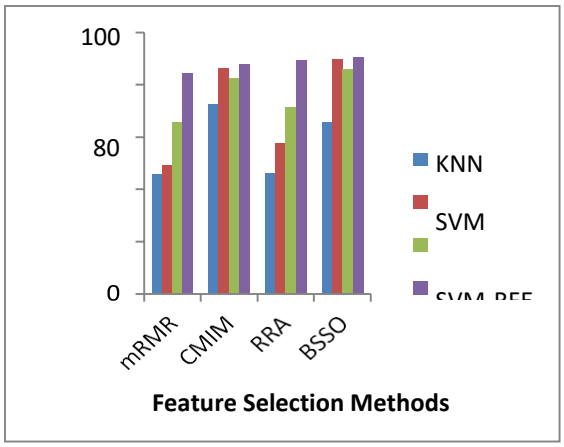

FIGURE 5. ACCURACY RESULTS OF FSMETHODS VS. CLASSIFIERS

Figure 5 proposed DAEDC classifier gives higher accuracy results for BSSO algorithm is 93.48%, similarly it also gives highest accuracy results for other FS methods such as mRMR, CMIM, and RRA methods like 89.75%,90.15%, and 91.71% respectively. The methods like KNN, SVM, SVM-RFE via BSSO algorithm also gives highest accuracy results of 88.18%,90.17%, and 91.36%
respectively(Refer Table 1).

5. CONCLUSION AND FUTURE WORK

The illness in brain, central nervous system and spinal cord leads to the rise in psychological problems. In this paper the work on feature selection (FS) algorithms is introduced to analysis the irrelevant and redundant features to increase the detection rate of neurological disorder. In this paper, a novel method is proposed for feature selection based on Binary Swallow Swarm Optimization (BSSO). The binarization of the continuous swallowswarm metaheuristics is carried out using a special function. Deep Auto- Encoder Based Data Clustering (DAEDC) algorithm is proposed in the data layer of a dataset with selected features is firstly mapped to the code layer, which is then used to reconstruct the data layer to classify the SNP data. The objective is minimizing the reconstruction error as well as the distance between data points and corresponding clusters in the code layer. The performance of classification illustrates that the selected SNPs have a meaningful impact on complex diseases or not. This advantage has decreased the time spent and the computational burden in the learning and operation phases. In conclusion, the proposed method was applied to SNP dataset, in which the precision, recall, F-measure, and accuracy were utilized to evaluate the performance of this method. The results displayed that the proposed BSSO based FS approach has succeeded to identify the significant SNPs in complex disease; the DAEDC algorithm was able to classify the healthy and patient samples with high accuracy according to the significant SNPs. In future the present system is applied to other complex diseases such as liver, breast

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 07-2022
DOI 10.17605/OSF.IO/AH3PB

cancer, Colorectal Cancer, Thyroid Cancer, and Autism.

REFERENCES

1. Kathiresan S, Voight BF, Purcell S, Musunuru K, Ardissino D, Mannucci PM, Anand S, EngertJC, Samani NJ, Schunker H, et al: Genome- wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nature genetics. 2009, 41 (3): 334-341. 10.1038/ng.327.
2. Burton PR, Clayton DG, Cardon LR, CraddockN, Deloukas P, Duncanson A, KwiatkowskiDP,McCarthy MI, Ouwehand WH, SamaniNJ, et al: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007, 447(7145): 661- 678. 10.1038/nature05911.
3. Abraham, G., and Inouye, M. (2015). Genomicrisk prediction of complex human disease and its clinical application. Curr. Opin. Genet.Dev. 33, 10–16. doi:10.1016/j.gde.2015.06.005
4. Abraham, G., Kowalczyk, A., Zobel, J., and Inouye, M. (2013). Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. Genet. Epidemiol. 37, 184– 195. doi: 10.1002/gepi.21698.
5. Ramirez-Gallego, S.; Mourino-Talin, H.; Martinez-Rego, D.; Bolon-Canedo, V.;Benitez, J.M.; Alonso-Betanzos, A.; Herrera,
F. An Information Theory-Based Feature Selection Framework for Big Data Under Apache Spark. IEEE Trans. Syst. Man Cybern. Syst. 2018, 48, 1441–1453.
6. Bolon-Canedo, V.; Sanchez-Marono, N.; Alonso-Betanzos, A. Feature Selection for High-Dimensional Data; Springer: Heidelberg, Germany, 2015; ISBN 978-3-319-21857-1.
7. Singh, D.; Singh, B. Hybridization of Feature Selection and Feature Weighting for High Dimensional Data. Appl. Intell. 2019, 49, 1580–1596.
8. Hamedmoghadam, H.; Jalili, M.; Yu, X. An Opinion Formation Based Binary OptimizationApproach for Feature Selection. Phys. A Stat. Mech. Its Appl. 2018, 491, 142–152.
9. Banitalebi, A.; Aziz, M.I.A.; Aziz, Z.A. A Self-Adaptive Binary Differential Evolution Algorithm for Large Scale Binary OptimizationProblems. Inf. Sci. 2016, 367, 487–511.
10. Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B. and Yang, G.Z., 2016. Deep learning for health informatics.
IEEE journal of biomedical and health informatics, 21(1), pp.4-21.
11. Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu,
M.S. Lew, Deep learning for visual understanding: A review, Neurocomputing 187(2016) 27–48, http://dx.doi.org/10.1016
/j.neucom. 2015.09.116.
12. Y. Bengio, Practical recommendations forgradient-based training of deep architectures, in: Neural Networks: Tricks of the Trade, Springer, 2012, pp. 437–478.
13. J.M. Alvarez, M. Salzmann, Learning thenumber of neurons in deep networks, in: Advances in Neural Information Processing Systems, 2016, pp. 2270–2278.
14. N. Srivastava, G.E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simpleway to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.
15. Wu, Q., Ye, Y., Liu, Y. and Ng, M.K., 2012. SNP selection and classification of genome-wide SNP data using stratified sampling randomforests. IEEE transactions on nanobioscience, 11(3), pp.216-227.
16. Nguyen, T.T., Huang, J.Z., Wu, Q., Nguyen,
T.T. and Li, M.J., 2015, Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. In BMC genomics (Vol. 16, No. 2, pp. 1-11). BioMed Central.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 07-2022
DOI 10.17605/OSF.IO/AH3PB

17. Kotlarz, K., Mielczarek, M., Suchocki, T., Czech, B., Guldbrandtsen, B. and Szyda, J.,2020. The application of deep learning for the classification of correct and incorrect SNP genotypes from whole-genome DNA sequencing pipelines. Journal of applied genetics, 61(4), pp.607-616.
18. Boutorh, A. and Guessoum, A, Complex diseases SNP selection and classification by hybrid Association Rule Mining and Artificial Neural Network—based Evolutionary Algorithms, Eng. Appl. Artif. Intell. 51 (2016) 58–70.
19. R. Alzubi, N. Ramzan, H. Alzoubi, A. Amira, Ahybrid Feature Selection Method for complex diseases SNPs, IEEE Access 6 (2018) 1292–1301.
20. Pirmoradi, S., Teshnehlab, M., Zarghami, N. and Sharifi, A., 2020. A Self-organizing Deep Auto-Encoder approach for Classification of Complex Diseases using SNP Genomics Data. Applied Soft Computing, 97, pp.1-12.
21. Sun, S., Miao, Z., Ratcliffe, B., Campbell, P., Pasch, B., El-Kassaby, Y.A., Balasundaram, B.and Chen, C., 2019. SNP variable selection by generalized graph domination. PLoS One,14(1),pp.1-18.
22. Potdar K., T.S. Pardawala, C.D. Pai, A comparative study of categorical variable encoding techniques for neural network

    classifiers, Int. J. Comput. Appl. 175 (4)(2017) 7–9.

23. Uppu, S., Krishna, A. and Gopalan, R.P., 2016.A deep learning approach to detect SNP interactions. J. Softw., 11(10), pp.965-975.

24. Neshat, M.; Sepidnam, G.; Sargolzaei, M. Swallow Swarm Optimization Algorithm: A New Method to Optimization. Neural Comput. Appl. 2013, 23, 429–454.

25. Bouzidi, S.; Riffi, M.E. Discrete Swallow Swarm Optimization Algorithm for Travelling Salesman Problem. In Proceedings of the ACMInternational Conference Proceeding Series, Rabat, Morocco, 21–23 July 2017; ACM Press:New York, NY, USA, 2017; Volume F1305, pp.80–84.

26. Song, C., Liu, F., Huang, Y., Wang, L. and Tan,T., 2013, Auto-encoder based data clustering. In Iberoamerican congress on pattern recognition (pp. 117-124). Springer, Berlin, Heidelberg.

27. Sakar, C.O., Kursun, O. and Gurgen, F., 2012. A feature selection method based on kernel canonical correlation analysis and the minimum Redundancy–Maximum Relevance filter method. Expert Systems withApplications, 39(3), pp.3432-3437.
28. Alzubi, R., Ramzan, N., Alzoubi, H. and Amira,A., 2017. A hybrid feature selection method forcomplex diseases SNPs. IEEE Access, 6,pp.1292-1301.