International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 21 (2019) pp. 4077-4084

Ensemble Machine Learning for Leukemia Cancer Diagnosis based on Microarray Datasets

Nashat Alrefai[1,2]

[1] *Basic & Applied Scientific Research Center, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, 31441, Dammam,Saudi Arabia.*

[2]*Department of mathematics, College of Science, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, 31441, Dammam, Saudi Arabia.*

Abstract

Background: Leukemia is defined as cancer of the body's blood-forming tissues, including the bone marrow and the lymphatic system. Nowadays, microarray gene expression datasets consider an essential source of data which is used in cancer classifications. However, due to the small size of samples compared to the high dimensionality of microarray data, many data mining techniques have failed to distinguish the most relevant and informatics genes. Therefore, combining several classifiers can improve the feature selection procedure and classification accuracy issues. The current study aims to propose a robust and accurate method for the leukemia disease diagnosis by utilizing an ensemble learning.

Methods: In this paper, Particle Swarm Optimization (PSO) method along with an ensemble learning method work together for feature selection. To clarify, the ensemble method used a combination of four classifiers: Support Vector Machine (SVM), k-Nearest Neighbor (kNN), Naïve Bayes (NB) and Decision Tree (C4.5) used to generate fitness function used in PSO as optimal solution to cover all the search space in shortest time with guaranty to choose the best number of meaningful genes that lead to improve the diagnosis of leukemia cancer. The current method applied on leukemia microarray gene expression, which is labeled and binary class (Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL)).

Results: The analysis demonstrates the usefulness of the proposed method from the side of the accuracy of classification on microarray datasets and the result of performance is superior than other individual classifiers.

Conclusions: The ensemble learning as a method of machine learning has the ability to be used for leukemia disease diagnosis and other diseases in the medical field with high accuracy.

Keywords: Ensemble Learning, cancer diagnosis, Particle Swarm Optimization, Gene Expression Microarray, Feature Selection.

1. INTRODUCTION

Cancer is one of the most common diseases with a high death rate among humans. It is considered the second reason of death in all continents of the world, in 2018 approximately 9.6 million deaths caused by cancer. In other words, about 1 in 6 deaths among humans are due to cancer. The most common types of cancer are lung (2.09 million cases), breast (2.09 million cases), colorectal (1.80 million cases) and prostate (1.28 million cases) [1]. With the early and accurate diagnosis of cancer, survival will increase from 56% to more than 86%, the Cancer death rate can be reduced in the case of early detected and treated [2]. Therefore, an accurate and reliable system is necessary for the early diagnosis of cancer.

Microarray dataset analysis and classification procedure have proved strongly that it provides an effective way for the effective diagnosis of diseases especially in cancers. Microarray device can be used to measure expression levels of a huge number of genes at the same time in a cell assortment, and lastly, the output of the microarray device is microarray data. Gene expression data is also another well-known name of microarray data [3]. A microarray dataset also is known as gene expression profile is usually constructed as a two dimensional array N×(M+1); where N is tissue samples represented as the number of rows or instances, and M is the gene expression level represented as the number of columns or features or attributes, one was added to the last column to present the class, usually, this class can be labeled, unlabeled or combination between them.

Cancer classification denotes to the procedure of model building on a dataset such as microarray gene expression datasets and then differentiating the value of the class for each instance in the sample, by this procedure the model was produced. Hence, the diagnosis results can aid doctors to follow suitable treatment protocol for the patients, particularly in the early time of disease diagnosis and treatment. Recently, many classification techniques were developed in the machine learning field and a considerable number of them were applied in cancer classification [4]. Nevertheless, some considerable difficulties would be clearly occurring according to the structured of microarray data when using the base learning algorithms. Moreover, there are some drawbacks during dealing with microarray data classification such as (1) Microarray ingrained holds a huge number of genes which called high dimensionality on the other side small number of samples which called low sample size, in machine learning this issue known as the curse of dimensionality which lead to higher risk of over-fitting, (2) The microarray data is related to an assortment of uncertainty due to the process of

microarray data acquiring, e.g. image processing, hybridization, fabrication, etc., always add various sources of noise which causes from the variation in the data that we cannot explain, and (3) Most of the genes are irrelevant to the classification of various tissue types [5] [6].

Feature selection is used to enhance cancer classification performance in microarray gene expression analysis, Feature selection essentially boosts to create better accuracy result whereas requiring fewer samples. It can be used to detect and remove undesirable, irrelevant and redundant features from data that do not contribute to the accuracy of a classification model.

The most challenge in microarray datasets is the high dimensionality, in this case, PSO, which was first suggested by Kennedy and Eberhart in 1995 [7], used for dimensionality reduction and to cover all the search space. Investigating the performance of the standard PSO for the classification of high dimensional data is high demand; Feature selection is an optimization issue where the objective is to choose the minimum number of features that have the maximum information. To apply PSO to the feature selection problem, it need first to map features selection/deselection using a representation suitable for PSO (usually continuous values representing the particle's position), develop the particles evaluation function, generate the initial swarm, and repeatedly apply the PSO steps of particle evaluation and update their velocity and positions till a predefined stopping criterion is met. There is a need to define some fitness function to compare the particles, the fitness function of the PSO can consider the F-score of the ensemble of the classifiers on the training set, this can reduce the dimensionality, avoid overfitting and improve class imbalance.

The Accuracy and diversity of individual classifier can guaranty better generalization capability through combining classifiers by using ensemble learning technique. And also, ensemble learning is considered a robustness option to solve class imbalance problem [8]. Reduce the dimensionality, avoid overfitting, improve class imbalance and concept drift problem all of these challenges can individual classifier be prone to it. In contrast, ensemble learning takes the benefit from the robustness of each classifier and minimize the error of individual one to enhance the model generalization ability to increase the accuracy and enhance the classification performance by Override most of the previous challenges.

Ensemble learning is defined as the use of algorithms and tools in machine learning and other areas, to produce a cooperative procedure, where multiple learners are more effective than an individual learner. Ensemble learning can be used in many fields such as disease diagnosis, finance, bioinformatics, healthcare, manufacturing, geography [9], for flexibility and enhanced results.

Human genes are very large in numbers. There are thousands or even millions of genes in human but only few of these genes have its own function while others have not been discovered its function yet. This also applies to the microarray data, this technology usually produces large datasets with thousands of genes expression values in a cell mixture, but the numbers of samples are very low. Since not all genes are relevant in determining certain type of disease, thus it is an important matter to select only informative genes out of all genes to provide enough information about a disease [10]. These genes will then be used to train the classifiers in order to construct rules to classify future unknown tissue samples into their appropriate classes. By selecting only informative genes can reduce the data dimensionality to be processed by the classifier, reducing the run time and improving the classification performance. Below are the reasons that contribute to the need of gene selection: reduce the common mistakes, removes the irrelevant genes or meaningless genes, reduce the dimension for search space, reduce a complex space and the run time, reduce clinical setting cost and improve classifiers performance.

Some classifiers such as SVM, k-NN and C4.5 can investigate it is robust to noise and outliers [11]. Moreover, according to previous study [12]. All these classifiers performed consistently well in microarray data, and also, they are from different classifications of algorithm, that mean they belong to different practices i.e. unstable, probabilistic and stable.

Every classifier has its own pros and cons and when selecting different types of classifiers, it can take advantage of all of them. Certainly, it depends on the way how to select in order to combine the result. Kotsiantis [13] adopt a classification algorithms comparison explained in Table1.

Table1: Classification Algorithms Comparison [13]

Classsifier	Decision Tree	Neural Networks	Naïve Bayes	kNN	SVM	Rule Learner
The accuracy	●●	●●●	●	●●	●●●●	●●
Training velocity	●●●	●	●●●●	●●●●	●	●●
Classification velocity	●●●●	●●●●	●●●●	●	●●●●	●●●●
missing values allowance	●●●	●	●●●●	●	●●	●●
Irrelevant attributes Tolerance	●●●	●	●●	●●	●●●●	●●
Tolerance to redundant attributes	●●	●●	●	●●	●●●	●●
Highly interdependent attributes Tolerance	●●	●●●	●	●	●●●	●●
Dealing with discrete/binary/continuous attributes	●●●●	●●●	●●●	●	●●	●●●
Noise Tolerance	●●	●●	●●●	●●●	●●	●
Dealing with overfitting	●●	●	●●●	●	●●	●●
Attempts for incremental learning	●●	●●●	●●●●	●●●●	●●	●
Explanation ability/Transparency	●●●●	●	●●●●	●●	●	●●●●
Model parameter handling	●●●	●	●●●●	●●●	●	●●●

Note: ●●●● denote the best, and ● denote the worst.

2. RELATED WORKS

We present some studies using data mining techniques that developed some methods for cancer disease diagnosis in Table 2. From this table, glance view can explore that most of the methods have been developed by a single classifier and there is no orientation for ensemble learning methods at that time. Moreover, the most methods developed by supervised, unsupervised and semi-supervised learning methods in the previous researches did not use ensembles learning for diseases diagnosis. As the diagnosis accuracy of standard supervised, unsupervised and semi-supervised learning methods can be improved by ensemble learning methods, recently ensemble learning offered promising results in medical field and diseases diagnosis especially when using microarray as a dataset, in this study, a new method is proposed using ensembles of four classifiers SVM, NB, C4.5 and k-NN, simultaneously work together with PSO which was used as a search method for dimensionality reduction and to cover all the search space. So the goal is Investigate the performance of the standard PSO for the classification of high dimensional data, Feature selection is an optimization issue aims is to choose the minimum number of features that have the maximum informative with guaranty to cover all the search space in minimum time comparing with traditional search methods such as greedy stepwise, best first and ranking search to enhance the predictive accuracy of the cancer disease diagnosis systems.

Table 2: Related Work on Leukemia Cancer Disease Diagnosis Based on Microarray Dataset

Disease	Authors	Techniques					
		C4.5	SVM	NB	kNN	RF	EL
Leukemia	Jinyan [14]	√					
	Arunkumar [15]					√	
	Bouazza [16]	√	√	√	√		
	Subhajit [17]				√		
	Mollaee [18]						√
	Mukesh [19]				√		
	Luis [20]		√				
	Chandra [21]		√	√			

Note: C4.5: Decision Tree, SVM: Support Vector Machine, NB: Naïve Bayes, kNN: k-nearest neighbors, RF: Random Forest, EL: Ensemble Learning.

3. METHODOLOGY

3.1 Research methodology

The generalization ability of an ensemble is often much stronger than that of base learners. The base classifiers used in ensemble learning which is proposed in this article are C4.5, SVM, NB and k-NN, under wrapper based feature selection model was design by hybridized PSO with the combination of the classifiers to decrease the dimensionality and increase the accuracy that leads to enhance the classification performance, there are four classifiers were used in the current model NB, SVM, C4.5 and k-NN, with PSO which was used as a search method for dimensionality reduction and to cover all the search space, but PSO need fitness function, the fitness function of the PSO in this model was the F-score of the classifiers on the training set, fitness = FScoretrain, However, by ensemble learning can generate one fitness value from classifiers combination and use it for PSO. After that aggregate all the models together and use majority vote technique to create the final target model. According to the related work in the previous section there are many supervised techniques proposed for dimensionality reduction space and cancer classification by using microarray analysis. In Figure1 the framework illustrates the phases of the proposed models.

PSO is used for feature subsets search and another classifier (C4.5, SVM, NB and k-NN) are used as a base classifier. Each classifier has its strength for example SVM is the highest classification accuracy and tolerant to irrelevant attribute comparing with other three classifiers, k-NN is the best according to the speed of the training and in incremental learning, NB is very strong in training and classification speed and allowance to missing values, C4.5 is very strong in dealing with discrete, binary and continuous attributes and also in classification speed.

3.2 PSO for Feature Selections

PSO was proposed by Kennedy and Eberhart [7], which is consider evolutionary computation technique for feature selection. Later, the inertia weight was introduced by Shi [22], to use in the particle swarm optimizer to output the standard algorithm of PSO. This PSO algorithm goes through several steps, the first step is initializing the population with random solution called particle. Every particle is process as a point denoted by x in an S-dimensional space. The ith particle is denoted by $X_i = (x_{i1},x_{i2},...,x_{iS})$. The best previous location '$p$best' of any particle is recorded and denoted by $P_i = (p_{i1},p_{i2},...,p_{iS})$, in which this location giving the best fitness value. '$g$best' represent the index of the best particle between all population of the particles. The change rate in the location (velocity) for particle i is denoted by $Vi = (v_{i1}, v_{i2},..., v_{iS})$. The particles are formed as the following formula:

$$v_{id} = w * v_{id} + c_1 * rand() * (p_{id} - x_{id}) + c_2 * Rand() * (p_{gd} - x_{id}) \quad (1)$$

$$X_{id} = x_{id} + v_{id} \quad (2)$$

Where the variable $d=1, 2, .. . ,S$. and w denote to the inertia weight; which is positive linear function that changing with time according to the repetition of generation. Appropriate chosen of the inertia weight offers a stability between local and global search, and this leads to less iteration on average to discover enough optimal solutions. In Eq. (1) the acceleration constants c1 and c2 represent the weighting of the random acceleration terms that move every particle toward $g$best and $p$best locations.

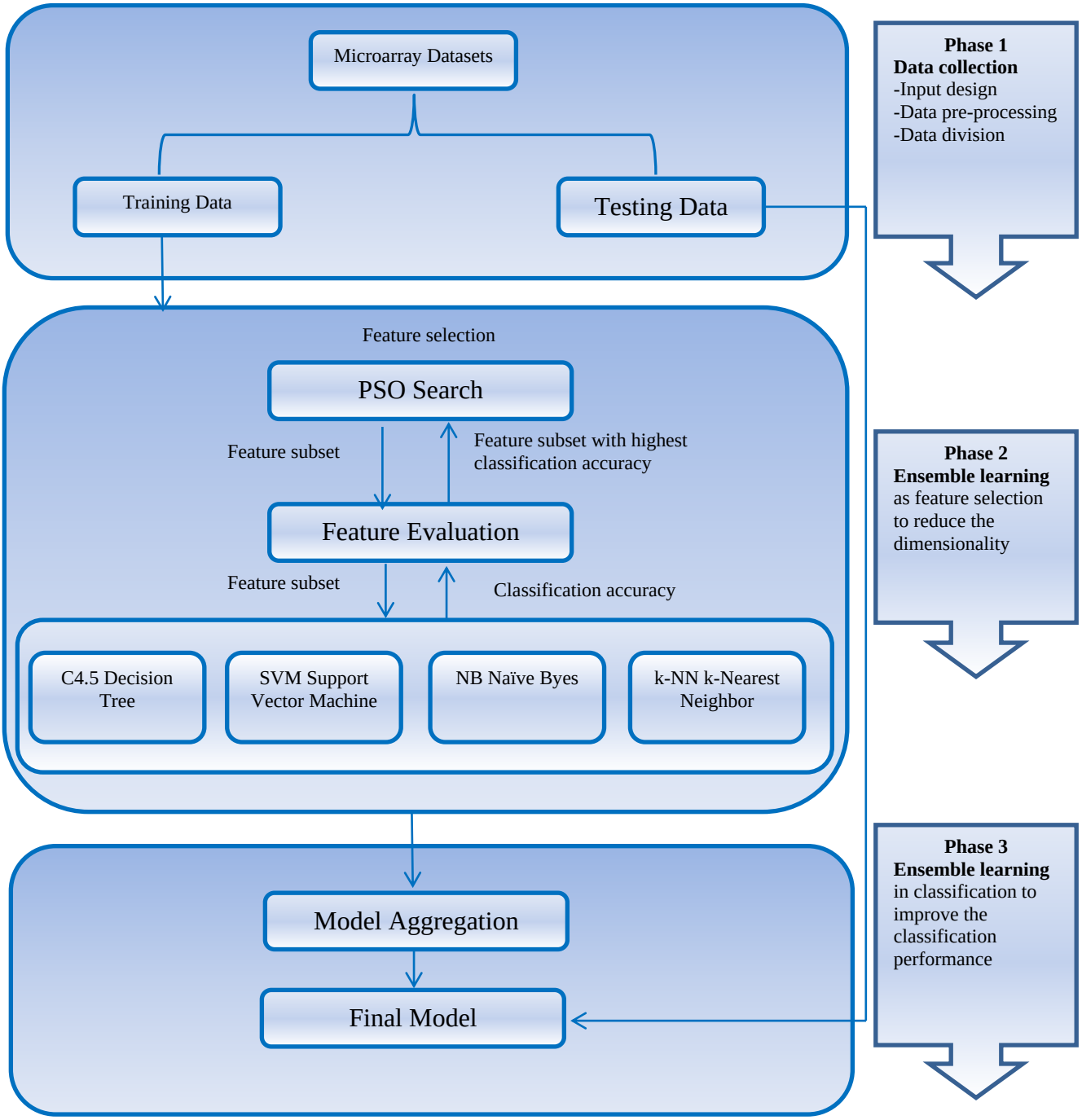

Figure 1: Research Methodology

When the value of the location is low, that leads the particles to go far from the target area; on the other hand, high values lead to a surprising movement forward, or past, target locations. *Rand()* and *rand()* are two different functions generate values between the range [0, 1]. Equation number (1) divided into three parts; the first part supplies the particle during flying with some ability of memory allowing the survey of new areas of search space. The second part is the knowledge part, which shows how the particle itself thinks in private way. The last part is a part related to the social behavior of the particle inside the population, which denotes the cooperation between the particles. The new velocity of the particle is measured in Eq.(1) according to its earlier velocity and the distances of its recent location from its personal best practices (position) and the group's best practices. Then the new location will be the target in which the particle flies toward, according to Eq.(2). The predefined fitness function used to measure the performance of each particle. The algorithm of the PSO procedure is given below.

PSO Algorithm

Inputs: *x*, is training data where $\sum_{x=1}^{n} x$ of X data from training dataset with *y* attributes

Begin

max_fitness ← y

for *i*=1 **to** *m*

particle$_i$ ← randomly initialize possible position. (1 feature is chosen, 0 otherwise)

particle$_{i_lbest}$ ← particle$_i$

end

while (curr_fitness < max_fitness) ***do***

Read data with respective feature subset (as represented by a particle) from input, X

for *i*=1 **to** *m* do

Evaluate fitness for particle$_i$ and particle$_{i_lbest}$ according to Equation (6)

if particle$_{i_fitness}$ > particle$_{i_lbest_fitness}$

then particle$_{lbest}$ = particle$_i$

particle$_{i_new_vel}$ update velocity according to Equation (1)

particle$_{i_new_pos}$ update location according to Equation (2)

end

G$_{best}$ ← best of (particle$_{1_lbest}$, particle$_{2_lbest}$, …, particle$_{m_lbest}$)

G$_{best_fitness}$ ← best of (particle$_{1_lbest_fitness}$, particle$_{2_lbest,_fitness}$ …, particle$_{m_lbest}$)

curr_fitness ← G$_{best_fitness}$

return G$_{best}$

End

Where, *Inertia weight+ Individual weight+ Social weight=1*, and the values for each of them computed by using three-parent mask-base crossover 3PMBCX instead of the notion velocity in order to determine the new position, and clearly defined in Table 3. The fitness function is usually defined as the accuracy of the classification using the features selected by every particle.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (5)$$

Fitness of the proposed feature(R) = majority voting (F1score-SVM, F1score-NB, F1score-C4.5, F1score-k-NN) (6)

Table 3. Parameters to Get Subset Selection and Classification

Default parameters	Value
Population size	20
Number of iteration	20
Report frequency	20
Mutation probability	0.01
Inertia weight	0.33
Individual weight	0.34
Social weight	0.33

4. ENSEMBLE EVALUATION

We classified leukemia patients using benchmark dataset which have binary class, the first value of the class is ((AML) Acute Myeloid Leukemia and the second value is (ALL) Acute Lymphoblastic Leukemia). Moreover, the WEKA tools were used to feature selection and classification. The microarray dataset used in this experiment had 7,129 attributes and the number of samples is 72, also divided into 38 sets as training sample data, and 34 sets as test sample data. The dataset rely upon in this article is the gene expression data available from the website http://eps.upo.es/bigs/datasets.html [23].

The first step after determining the dataset is data pre-processing The following steps were used to pre-process the data to make it clean and suitable for prediction modeling: replace the missing values (All missing values in a dataset will replace with the means and modes from the training data), remove the outlier and extreme values, attributes that have the same value for more than 99% of the patients are removed.

We used the proposed ensemble method which ensemble the four classifiers together, by ensemble learning can generate one fitness value from classifiers combination and use it for PSO as fitness function used together to find the optimal number of features which consider sufficient to train the classifier to improve the accuracy and enhance the model performance. The results in Table 4 compare the performance of each individual classifier SVM, NB, C4.5 and k-NN. After that measure the performance of each classifier wrapper with PSO and finally find the accuracy of the proposed ensemble-method.

Table 4: Shows the Results Obtained for Leukemia Dataset

		Accuracy %	Precision	Recall	F-Measure	ROC Area	# of selected features
Base classifiers	Decision Tree (J48)	84.2105 %	0.842	0.842	0.842	0.808	7130
	Support Vector Machine (SVM)	94.7368 %	0.951	0.947	0.946	0.909	7130
	Naive Bayes (NB)	94.7368 %	0.951	0.947	0.946	0.922	7130
	k-nearest neighbor (k-NN)	89.4737 %	0.908	0.895	0.887	0.845	7130
Wrapper (classifier with PSO)	C4.5-PSO	97.3684 %	0.976	0.974	0.974	0.981	1231
	SVM-PSO	97.3684 %	0.975	0.974	0.973	0.955	1615
	NB-PSO	100 %	1.000	1.000	1.000	1.000	1512
	k-NN-PSO	92.1053 %	0.920	0.921	0.920	0.889	1604
Ensemble (classifiers) with PSO	Ensemble (C4.5, SVM, NB, K-NN)-PSO	**100 %**	**1.000**	**1.000**	**1.000**	**1.000**	**1629**

5 DESCUSSION

From Table 4 the accuracy in percentage for our proposed method is 100.00 % and the number of selected attributes are 1629 which consider the most relevant and informatics genes, the time which taken to build the model is 0.02 seconds and the cross validation with 10-folds are used to validate the results, the accuracy is obtained after applying majority voting which is ensemble method to the wrapper of PSO and the four selected classifiers, if half or more of the classifiers vote to particular percentage then it was selected in each instant, in our case indicate that for each patient half or more than half of classifiers vote to the correct prediction, so it was selected, In addition, the proposed model was more stable and the result can be generalize, also the model was tested by using average probabilities and give the same result, but another combination rule such as product of probabilities, minimum probabilities, maximum probabilities and median are not suitable for our case.

The proposed method selects just 1629 features from 7129 features which show the robustness of this method comparing with other methods, thanks to PSO which cover the search space and help the classifiers to eliminate the irrelevant genes, this prove that the method can select the most informatics, relevant and non-redundant attributes. The percentages of attributes which are selected are illustrated in Figure 2.

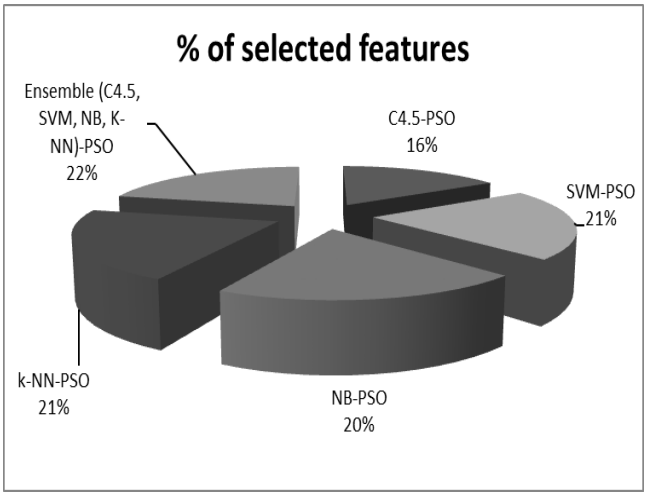

Figure 2: The Percentages of Selected Features

The evaluation method used in this paper is depending on the confusion matrix. The confusion matrix consider a visualization tool usually used to present the classifiers accuracy during classification procedure [24]. It is used to present the relationships between actual and predicted classes. The level of efficiency of the classification model is calculated with the number of accurate and inaccurate classifications in each possible value of the instance being classified in the confusion matrix. The confusion matrix is two dimensional array contains the negative and positive

values of actual and predicted class for instances, the four categories illustrate in Table 5.

Table 5: The Confusion Matrix

		Predicted	
		positives	negatives
Actual	Positive	Number of true positive classes (TP) 27	Number of false negative classes (FN) 0
	negative	Number of false positive classes (FP) 0	Number of true negative classes (TN) 11

True Positives (TP) represent the correctly predicted as positive values of instances, True Negatives (TN) represent the correctly predicted as negative values of instances, False Positives (FP) represent the incorrectly predicted as negative values of instances and False Negatives (FN) represent the incorrectly predicted as positive values of instances. The accuracy for each method was presented in Figure 3.

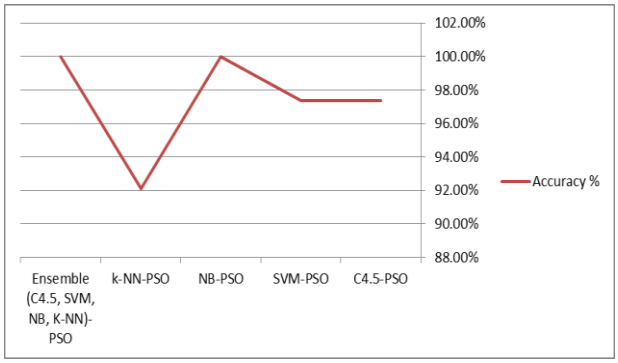

Figure 3: The Accuracy of The Selected Methods

6 CONCLUSION

A new ensemble genes selection method was proposed in this paper, where ensemble of most common classifiers combine together and work as fitness function to PSO to find the optimal number of informatics genes to improve the classification accuracy. As a result, the obtained gene subset has the most information and has capability for classification. Moreover, utilized from the diversity of classifiers rather than individual base classifier which give robustness and enhance the performance of prediction after integrated the output by the majority voting.

This work has some limitations, firstly, the experimental results on leukemia cancer public microarray dataset which is labeled and binary class contains two values (ALL, AML) have verified that our ensemble method outperforms others, But, computational time of our method is higher than other classifiers, but more stable for overfitting and the concept of drift. Secondly, because the size of microarray datasets is larg, so the training procedure need super computers or distributed system. Finally, the number of instances in microarray datasets is very low comparing with the high dimensionality; this is increase the hypotheses in the search space and makes the classifiers work difficult in decision making. As a result, semi supervised learning was suggested as future work to enlarge the size of the samples in the dataset.

REFRENCES

[1] "Cancer," *World Health Organization*. 22-Nov-2018.

[2] M. Montazeri, "Machine learning models in breast cancer survival prediction," *Technol. Heal. Care*, vol. 24, no. 1, pp. 31–42, Jan. 2016.

[3] Y. Peng, "A novel ensemble machine learning for robust microarray data classification," *Comput. Biol. Med.*, vol. 36, no. 6, pp. 553–573, Jun. 2006.

[4] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, Jan. 2015.

[5] X. Wang, M. J. Hessner, Y. Wu, N. Pati, and S. Ghosh, "Quantitative quality control in microarray experiments and the application in data filtering, normalization and false positive rate prediction," *Bioinformatics*, vol. 19, no. 11, pp. 1341–1347, Jul. 2003.

[6] M. S. Mohamad, S. Omatu, M. Yoshioka, and S. Deris, "An approach using hybrid methods to select informative genes from microarray data for cancer classification," *Proc. - 2nd Asia Int. Conf. Model. Simulation, AMS 2008*, pp. 603–608, 2008.

[7] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science*, 1995, pp. 39–43.

[8] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, "Classification with class imbalance problem: A review," *Int. J. Adv. Soft Comput. its Appl.*, vol. 7, no. 3, pp. 176–204, 2007.

[9] L. Rokach, "Ensemble Methods for Classification," *Data Min. Knowl. Discov. Handb.*, pp. 957–980, 2005.

[10] R. Maglietta *et al.*, "Selection of relevant genes in cancer diagnosis based on their prediction accuracy," *Artif. Intell. Med.*, vol. 40, no. 1, pp. 29–44, May 2007.

[11] G. H. Jowkar and E. G. Mansoori, "Perceptron ensemble of graph-based positive-unlabeled learning

for disease gene identification," *Comput. Biol. Chem.*, vol. 64, pp. 263–270, 2016.

[12] M. Morovvat and A. Osareh, "An Ensemble of Filters and Wrappers for Microarray Data Classification," *Mach. Learn. Appl. An Int. J.*, vol. 3, no. 2, pp. 01-17, 2016.

[13] S. B. Kotsiantis, "Supervised Machine Learning : A Review of Classification Techniques," vol. 31, pp. 249–268, 2007.

[14] J. Li, H. Liu, and N. Li, "Diagnostic rules induced by an ensemble method for childhood leukemia," *Proc. - BIBE 2005 5th IEEE Symp. Bioinforma. Bioeng.*, vol. 2005, pp. 246–249, 2005.

[15] C. Arunkumar and S. Ramakrishnan, "Attribute selection using fuzzy roughset based customized similarity measure for lung cancer microarray gene expression data," *Futur. Comput. Informatics J.*, vol. 3, no. 1, pp. 131–142, Jun. 2018.

[16] S. H. Bouazza, K. Auhmani, A. Zeroual, and N. Hamdi, "Selecting significant marker genes from microarray data by filter approach for cancer diagnosis," *Procedia Comput. Sci.*, vol. 127, pp. 300–309, Jan. 2018.

[17] S. Kar, K. Das Sharma, and M. Maitra, "Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique," *Expert Syst. Appl.*, vol. 42, no. 1, pp. 612–627, Jan. 2015.

[18] M. Mollaee and M. H. Moattar, "A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification," *Biocybern. Biomed. Eng.*, vol. 36, no. 3, pp. 521–529, Jan. 2016.

[19] M. Kumar, N. K. Rath, and S. K. Rath, "Analysis of microarray leukemia data using an efficient MapReduce-based K-nearest-neighbor classifier," *J. Biomed. Inform.*, vol. 60, pp. 395–409, Apr. 2016.

[20] L. H. S. Vogado, R. M. S. Veras, F. H. D. Araujo, R. R. V. Silva, and K. R. T. Aires, "Leukemia diagnosis in blood slides using transfer learning in CNNs and SVM for classification," *Eng. Appl. Artif. Intell.*, vol. 72, pp. 415–422, Jun. 2018.

[21] B. Chandra and M. Gupta, "An efficient statistical feature selection approach for classification of gene expression data," *J. Biomed. Inform.*, vol. 44, no. 4, pp. 529–535, Aug. 2011.

[22] Shi. Eberhart, "A modified particle swarm optimizer," in *1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98TH8360)*, 1998, pp. 69–73.

[23] "BioInformatics Group Seville." [Online]. Available: http://eps.upo.es/bigs/datasets.html. [Accessed: 13-Dec-2018].

[24] J. Han, M. Kamber, and J. Pei, "Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)," 2011.