SVM-Kmeans: Support Vector Machine based on Kmeans Clustering for Breast Cancer Diagnosis

Walaa Gad
Faculty of Computers and Information Sciences
Ain Shams University
Cairo, Egypt
Email: walaagad [AT] cis.asu.edu.eg

Abstract—**Breast cancer is the most common cancer in women, and is considered one of the most common causes of death. It increases by an alarming rate globally. Earlier detection and diagnosis could save lives and improve quality of life. In this paper, a new method for breast cancer diagnosis is presented. The proposed method, SVM-Kmeans, combines Kmeans, an unsupervised learning clustering technique, with Support Vector Machine (SVM), a supervised learning classifier. SVM-Kmeans determines number of clusters which achieves the best performance. Moreover, SVM-Kmeans removes irrelevant features using Chi-square feature selection method. This step speeds up SVM-Kmeans and solves curse dimensionality problem.**

We use Precision, Recall, and Accuracy performance measures to evaluate SVM-Kmeans using two breast cancer datasets. Experimental results show that SVM-Kmeans has a competitive performance compared to other methods in literature. Results show an accuracy rate achievement of 99.8%.

Keywords-component; Support Vector Machine; Kmeans; Breast cancer diagnosis; Classification

I. INTRODUCTION

The World Cancer Report [1] states that breast cancer is the second cause of women death after lung cancer. Globally, it increases at an alarming rate. In many countries, breast cancer has increased during the 20th century due to global changing and regional increasing in mammography [2].

Early and accurate diagnosis of breast cancer disease can lead to successful treatment. Analyzing cancer diagnoses help medical experts to predict breast cancer in a new patient. Therefore, machine learning methods are employed to improve and enhance the outcomes of existing methods that may help in disease diagnosis.

Many supervised learning classifiers [3]–[11] are introduced to help in disease diagnosis. Classifiers may be single, such as Naive Bayes (NB), Multilayer Perceptron (MLP), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). Multiple classifiers can be combined together to enhance the accuracy.

In this paper, a novel method, SVM-KMeans, is proposed for breast cancer diagnosis. The proposed method uses Support Vector Machine (SVM) classifier in conjunction with Kmeans clustering algorithm. Although, SVM provides good results in classification, but still needs more enhancement especially in disease diagnosis. SVM classifier has the ability to deal with very high dimensional data, and from computation perspective, SVM provides a fast training process [12].

In SVM-Kmeans, the clustering algorithm preserves the structure of the original dataset, and number of clusters is added to the training process. In addition, kernel and penalty factor parameters of SVM are defined as well. In clustering step, number of clusters *k* is usually defined by a domain expert. The proposed method aims at determining the number of clusters *k.* The unnecessary and irrelevant features are removed to speed up the computation time. Chi-square method, feature selection [13], is adopted to select the most important features.

In the proposed method, learning process consists of training and testing steps. In the learning process, SVM-Kmeans distributes data into *k* equally sized partitions. Training is applied on *(k-1)* folds. Then, testing is done on the remaining fold. For each fold, this process is repeated, and the average error rate is calculated to form *k*-fold estimate [14].

The proposed approach is evaluated using two datasets for breast cancer: Breast Cancer Wisconsin Diagnostic dataset (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC) [15], obtained from UCI machine learning repository. Experiments are conducted using 10-fold cross validation method. The obtained results are very promising where the accuracy approaches to 100% in case of the 10-fold method using WDBC, and 98% using WPBC, respectively. In addition, we compared the proposed method with the previously proposed methods in [3]–[6] using different measures: Precision, Recall, True Positive (Tp), False Positive (Fp) and Accuracy.

The paper is organized as follows. Next section presents an overview of the related work. The proposed SVM-Kmeans algorithm is described in Section III. Section IV shows the datasets and the experimental results. Finally, conclusions are drawn in section V.

II. RELATED WORK

This section summarizes some of the breast cancer diagnosis work found in literature. In [16], authors analyzed the performance of supervised learning algorithms for breast cancer diagnosis. Two datasets are used, Wisconsin Diagnostic Breast Cancer dataset (WDBC), and Breast Tissue dataset.

They compared the results among different algorithms such as SVM, Gaussian RBF kernel, Naive Bayes, RBF neural networks, J48, Decision trees, and simple CART. Accuracy, Precision, Recall and sensitivity measures are used to evaluate classifiers performance. Their experimental results showed that SVM-RBF kernel reached the highest accuracy 96.84% in WBC and 99% in Breast tissue.

In [17], a comparison is done between different classification algorithms, support Vector Machine (SVM) classifiers gives the best results using WDBC dataset for breast cancer.

Lavanya el al. [3] introduced feature selection method for classification to eliminate irrelevant attributes and increase classifier accuracy. They used filter, wrapper and hybrid approaches for feature selection. They compared among different classifiers using feature extraction compared to the same classifiers without feature selection. They concluded that the Decision tree classifier-CART had higher accuracy when applied on WBC and WDBC datasets. Without feature selection, CART algorithm showed accuracy 69.23%, 94.84%, and 92.97 using Breast Cancer, WBC, WDBC datasets respectively. They used two different methods for feature selection. Using principal component method, accuracies scored 70.63%, 96.99% and 92.09% in Breast Cancer dataset, WBC dataset and WDBC dataset respectively. Using Chi Squared method, accuracy reached 69.23% in Breast Cancer dataset, 94.56% in WBC dataset and 92.61% in WDBC dataset.

A parallel approach using neural network technique was proposed in [18].The input to a classifier is a training dataset. Each record is described by its attribute values and class label. Attributes could be discrete or numeric values. The goal is to induce a model or description for each class in terms of the attributes. Breast cancer database is used to train the neural network. In parallel approach, feed forward neural network model and back propagation learning algorithm, momentum and variable learning rate are used. The experiment is conducted by considering the single and multilayer neural network models. Results showed that 92% of test data were correctly classified and 8% were misclassified.

In [4], a hybrid approach was introduced. CART classifier with feature selection and bagging technique combined together. Bagging [19] is an ensemble method to classify the data with high accuracy. First, the decision trees are derived by building the base classifiers c_1, c_2,.., c_n on dataset samples. The final model or decision tree is derived as a combination of all base classifiers. In the hybrid approach, authors combined feature selection method, bagging and cart decision tree algorithm. Accuracy reached to 97% using Breast Cancer Wisconsin (Original) and 95.96% Breast Cancer Wisconsin (Diagnostic).

Gouda el al. [5] introduced a comparison between different classifiers. They used three different datasets, Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC). Accuracy and fusion matrix were calculated to compare between classifiers. The comparison was among Naive Bayes (NB), The Multilayer Perceptron (MLP), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision tree J48 classifiers. SVM recorded the highest accuracy value 96.9% compare to 95.5%, 95.2%,95.1%, and 94% in NB, MLP, J48, KNN respectively using WBC. SVM classifier was superior too using WDBC dataset. It recorded accuracy value 97.7% compared to 96.6%, 95.5%, 93.1%, 92.9% using MLP, KNN, J48 and NB respectively. Moreover, they combined more than one classifier together: (SVM, KNN, NB), (SVM, KNN and MLP) and (J48, NB, KNN). The best accuracy performance was 97.7% using SVM, KNN, NB and J48 multi-classifier with WDBC dataset.

III. THE PROPOSED METHOD SVM-KMEANS

The proposed SVM-Kmeans approach is a combination of clustering, feature selection and classification methods. Kmeans partitions data into *k* clusters and maintains the main distributions of the dataset. Then, the important features are selected using Chi-square to reduce the large number of features. In the last step, SVM is applied. Figure 1 lists the proposed SVM-Kmeans in a flowchart. The proposed model consists of:

- Read breast cancer dataset.
- Preprocess dataset.
- Partition dataset into k clusters.
- Select important features using Chi-square method.
- Build SVM classifier.
- Select the best performance parameters.

A. *Preprocessing*

In classification step using SVM, dataset features should be in real number format. Therefore, the preprocessing step transforms categorical features into numerical data. Then, normalization function [20] is performed.

$$F_{Normalization} = \frac{F - F_{min}}{F_{max} - F_{min}} \quad (1)$$

Learning process is divided into two steps training and testing. We use *k*-fold cross validation, the proposed method divides the dataset into *k* folds of equal sizes. The proposed method repeats this step for fold, and the average error is calculated to form the *k*-fold estimate [14].

B. *Clustering*

K-means is an important clustering algorithm in machine learning and pattern recognition [21]. It aims at partitioning a given dataset into a certain number of clusters. It selects centroids randomly, and assigns data points to the nearest centroids to minimize the inter-cluster similarity.

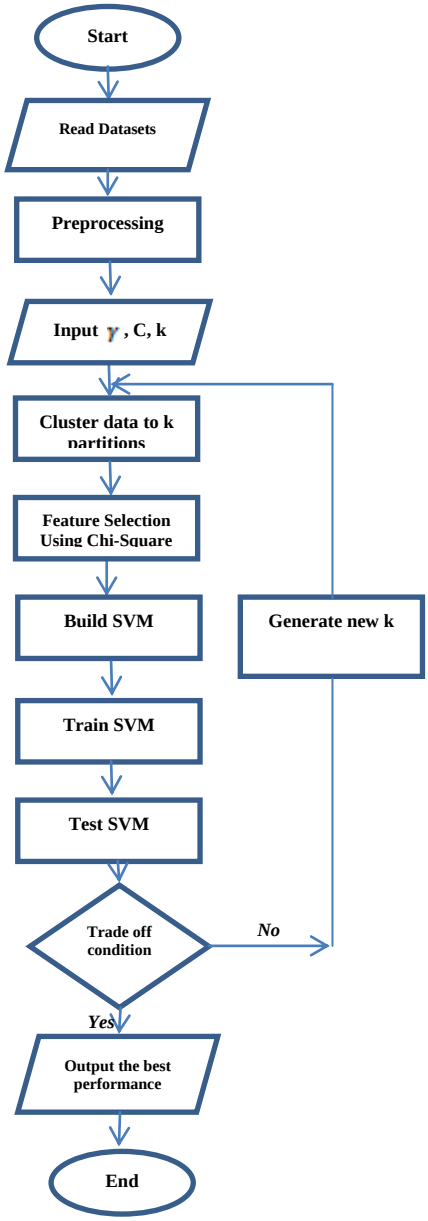

Fig. 1: Proposed SVM-Kmeans model.

Assume, there are n objects, $O_1,O_2,O_3,\ldots,O_n$. Each object is a d-dimensional vector. Kmeans aims to partition the n objects into k sets $S = S_1,S_{\ldots,} S_k$ so as to minimize the within-cluster sum of squares, which is the sum of distance functions of each point in the cluster to the k center. Kmeans aims to minimize the objectives function, which is defined as:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{O}\in S_i} \|\mathbf{O}-\boldsymbol{\mu}_i\|^2 \qquad (2)$$

where u_i is the mean of points in S_i.

C. Feature Selection

In feature selection step, important features are selected from original features using different techniques. Then, each feature is evaluated to determine its relevancy towards the classification using the measures: distance, dependency, information, consistency, classifier error rate.

In the proposed model, Chi-square is adopted to

- Reduce training time.
- Reduce classification over fitting.
- Remove irrelevant features to solve dimensionality problem.

Chi-Square (X^2) is a statistical method to test independence between two features. Chi-Square is defined as:

$$X^2(t,c) = \sum_{e_t\in 0,1} \sum_{e_c\in 0,1} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}} \qquad (3)$$

where t is a feature in class c, N is the observed frequency, E the expected frequency. e_t equals 1 if the object contains a feature t and e_t equals 0 if the object does not contain t. e_c equals 1 if the object is in class c and e_c equals 0 if the object is not in class c.

D. SVM-Kmeans Classifier

SVM algorithm is established by Vapnik [22]. SVM aims at maximizing the margin and the kernel trick to reach accuracy, and overcomes the problem curse of dimensionality. In classification, SVM solves the quadratic optimization problem in equation 4.

$$min \|w\|^2 + C \sum_{i=1..l} \xi_i \qquad (4)$$

In non-linear problems, many kernel functions are used. The Gaussian radius basis function (RBF), polynomial function, and the sigmoid function are the most popular kernel functions. SVM is defined as

$$f(x) = sgn\left(\sum_{i=1..n} \alpha_i K(x, x_i) + b\right) \qquad (5)$$

where x_i is a support vector, α_i is the coefficient of the support vector, n is the number of support vectors, b is the bias, K is the kernel function, and sgn is the sign function. The proposed model adopts the Gaussian RBF kernel in equation 6.

$$K(x_1, x_2) = exp(-\gamma \|(x_1 - x_2)\|^2) \qquad (6)$$

Defining the input parameters in supervised learning is called model selection. Some model selection methods have been proposed in [14] and [23]. Similar to [5], we set the same values to SVM parameters. The penalty factor C, kernel parameter, ζ values are 1, 0.01, 1.0E-12 respectively. We set

number of clusters *k* to 2, 4, 5, 6, and 8. The termination condition is verified by experiments to find the best performance with different *k* values.

Kmeans preserves the structure and distribution of the original data. Therefore, Kmeans is used in conjunction with SVM to find the best way to classify the dataset. Figure 2 shows the main idea of the proposed algorithm. Assume there is a set of positive and negative points. They are clustered using kmeans and compressed using clustering techniques to 5 clusters, C_1, C_2,.., C_5. Experiments showed that Kmeans preserves the same distribution and structure of the original data by these five cluster centers.

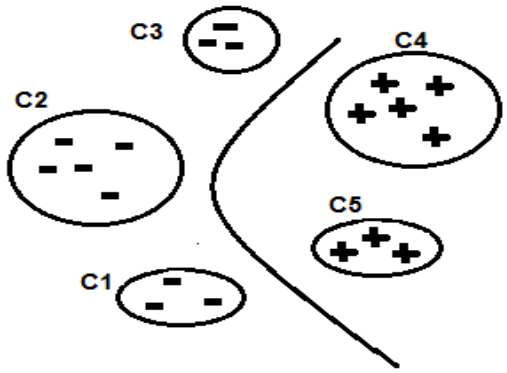

Fig. 2: SVM classifier based on Kmeans clustering.

IV. EXPERIMENTS AND RESULTS

A. Dataset

The proposed algorithm SVM-Kmeans is evaluated using two data sets: Breast Cancer Wisconsin Diagnostic dataset (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC) [15]. We obtained the two datasets from UCI machine learning repository.

Breast cancer features are extracted using fine needle aspirate (FNA) of breast mass images. Features describe the characteristics of the cell nuclei in the image [24]. Those features consist of ID, Diagnosis label and ten real-valued features [15].

Table I shows WDBC features in details. It consists of 569 instances, 32 features. Class distribution is 357 benign and 212 malignant. The main objective of this work is to predict the possibility of breast cancer occurrence in new patients by studying their features. Table II shows number of features, number of cases and number of class labels for each breast cancer dataset.

The proposed model replaces the missing values by appropriate values which is the corresponding mean value of the attribute. All attributes are represented in real valued measurement.

TABLE I: Breast Cancer Wisconsin dataset description

ID number	patient identification number
Diagnosis	M = malignant, B = benign
Ten real-valued features are computed for each cell nucleus	
Radius	mean of distances from center to points on the perimeter
Texture	standard deviation of gray-scale values
Perimeter	perimeter of the cell nucleus
Area	area of the cell nucleus
Smoothness	local variation in radius lengths
Compactness	perimeter^2 area - 1.0
Concavity	severity of concave portions of the contour
Concave	points number of concave portions of the contour
Symmetry	symmetry of the cell nuclei
Dimension	coastline approximation – 1

TABLE II: Breast cancer datasets description.

Dataset	# features	#patients	# Classes
WDBC	32	569	2
WPBC	34	198	2

B. Evaluation Measures

The following measures have been used: Recall, Precision, Tp Rate, Fp Rate, and Accuracy to evaluate the proposed model. The performance measures are defined as:

$$Precision = \frac{Tp}{Tp + Fp} \quad (7)$$

$$Recall = \frac{Tp}{Tp + Fn} \quad (8)$$

$$Accuracy = \frac{Tp + Tn}{Tp + Fp + Tn + Fn} \quad (9)$$

where:

- Tp is the number of cases correctly belongs to its class label.
- Tn is the number of cases incorrectly labeled as belonging to the class label.
- Fp is the number of cases incorrectly rejected from its class label.

- Fn is the number of cases, which are not labeled as belonging to the positive class label but should have been.
- The accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined.

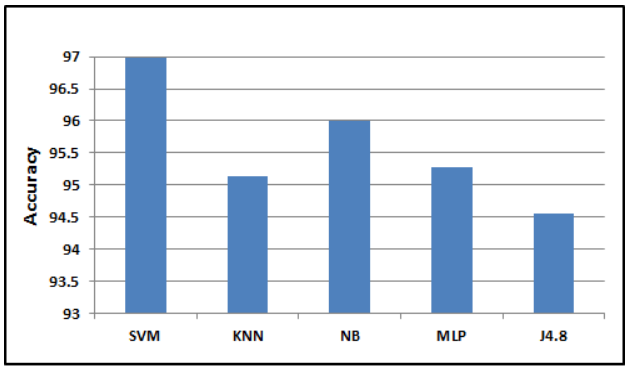

Fig. 3: Accuracy of different classifiers using WDBC.

In this paper, different classifiers are evaluated including MLP, NB, KNN, J4.8, SVM and others. Such comparisons are done to select the best classifier for breast cancer diagnosis. The Multilayer Perceptron (MLP) is based on neural network technique with three different layers input, output and hidden layers. Weighting coefficients are adjusted to find the most powerful output at the output layer. Naive Bayes (NB) classifier is a probabilistic classifier which is based on applying Bayes theorem independence assumptions between the features. K-Nearest Neighbor (KNN) is a type of lazy classifiers, where objects are classified according to their similarity to the k closest training examples in the feature space. J48 classifier is a decision tree which considers that each attribute of the data can be used to make a decision by splitting the data into smaller subsets.

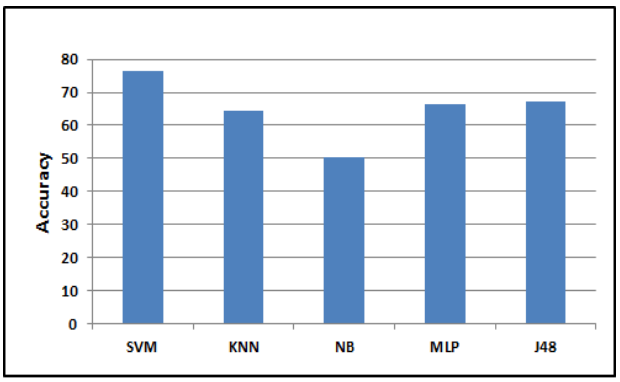

Fig. 4: Accuracy of different classifiers WPBC.

Figures 3 and 4 show the performance of SVM classifiers compared to KNN, MLP, J4.8 classifiers. The SVM performance outperforms the other classifiers. The results are consistent with the published ones in [5]. The SVM classifier reaches an accuracy of 96.8% and 76.2% using WDBC and WPBC datasets. Therefore, SVM is used in this work as a classifier.

In figures 5 and 6, SVM-Kmeans performance is shown versus number of clusters using precision, recall and accuracy measures for WDBC and WPBC datasets. The SVM-Kmeans reaches superior accuracy when number of clusters is 2. It achieves an accuracy of 99.9% and 98.9% for WDBC and WPBC respectively compared to different values of *k* starting from 2 till 10.

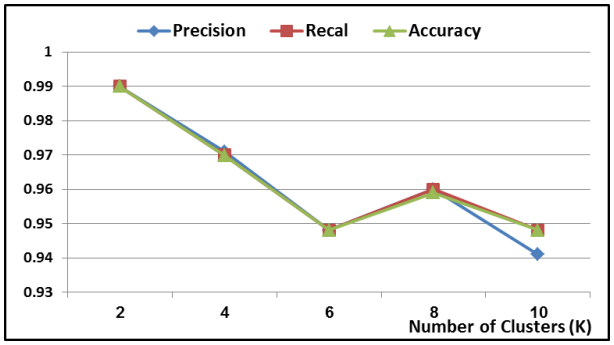

Fig. 5: Precision, Recall, Accuracy versus number of clusters (*k*) using WDBC dataset.

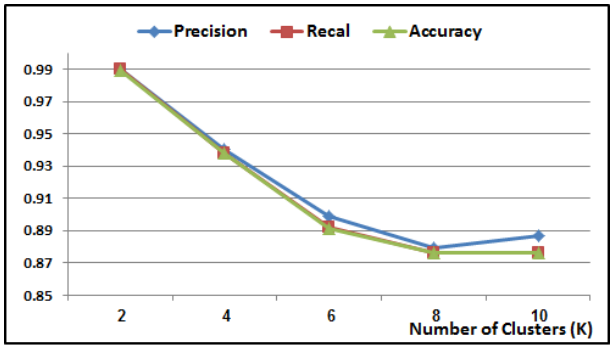

Fig. 6: Precision, Recall, Accuracy versus number of clusters (*k*) using WPBC

Table III lists comparisons among the proposed SVM-Kmeans and other methods found in the literature [3]–[6]. The SVM-Kmeans enhances the accuracy performance by 7.2%, 3.8%, 4.3% and 4.75 compared to methods in [3], [4], [7], [25] using WDBC dataset. Using WPBC, its performance outperforms methods in [5], [10], [11], where performance increases by 21.19%, 27.75%, and 22.18% respectively. Experimental results prove the efficiency of SVM-Kmeans to be employed to help specialists in the diagnosis of breast cancer, and to provide them with information that may help in making decision for disease diagnosis and saving new patients.

TABLE III: Results of SVM-Kmeans Accuracy compared with other approaches.

Method Reference	Method	Dataset	Accuracy
[3]	Cart	WDBC	92.61%
[4]	Hybrid Approach	WDBC	95.96%
[5]	SMO	WDBC	97%
[5]	SMO	WPBC	77.31%
[6]	CatfishBPSO	WDBC	98%
[7]	Fuzzy Rule Classification	WDBC	96.08%
[8]	Supervised Fuzzy Clustering	WDBC	95.57%
[9]	CBRGenetic	WDBC	97.37%
[10]	Jordan Elman Neural Network	WPBC	70.72%
[11]	RBF-SVM	WPBC	76.32%
[25]	Neuron-Fuzzy	WDBC	95.05%
Proposed	SVM-Kmeans	WDBC	99.8%
Proposed	SVM-SVM	WPDC	98.5%

V. CONCLUSION

The widespread of breast cancer among women with different races, in different countries, triggers an alarm to increase the attention on the prevention or early diagnosis of the disease. Early and accurate diagnosis can lead to successful treatment, and improve quality of life. Using computer science, especially data mining techniques, would significantly help in medical diagnosis. In this paper, we proposed a novel algorithm SVM-Kmeans. For breast cancer diagnosis, Support Vector Machine and Kmeans are deployed to predict the possibility of occurrence of breast cancer in a new patient. Moreover, feature selection is used to remove irrelevant features and solve the problem of curse dimensionality. SVM-Kmeans is applied on two most common datasets for breast cancer, together with many other classifiers proposed in literature. To evaluate the proposed algorithm, accuracy, precision, and recall performance measures are used. Experiments show that SVM-Kmeans has superior results compared to other approaches. It reaches up to 99.8% accuracy, 0.99 recall, and 0.99 precision.

REFERENCES

[1] Us cancer statistics working group. united states cancer statistics, 1999-2008 incidence and mortality web-based report. atlanta (ga): Department of health and human services, centers for disease control and prevention, and national cancer institute, 2012.

[2] M. Tanter, J. Bercoff, A. Athanasiou, T. Deffieux, J.-L. Gennisson, G. Montaldo, M. Muller, A. Tardivon, and M. Fink, "Quantitative assessment of breast lesion viscoelasticity: initial clinical results using supersonic shear imaging," Ultrasound in medicine & biology, vol. 34, no. 9, pp. 1373–1386, 2008.

[3] D. Lavanya and D. K. U. Rani, "Analysis of feature selection with classification: Breast cancer datasets," pp. 756–763, 2011.

[4] D. Lavanya and K. U. Rani, "Ensemble decision making system for breast cancer data," International Journal of Computer Applications, vol. 51, no. 17, pp. 19–23, August 2012, full text available.

[5] M. A. Gouda I. Salama and M. A. elghany Zeid, "Breast cancer diagnosis on three different datasets using multi-classifiers," pp. 36–43, 2014.

[6] L. Chuang, S. Tsai, and C. Yang, "Improved binary particle swarm optimization using catfish effect for feature selection," Expert Syst. Appl., vol. 38, no. 10, pp. 12 699–12 707, 2011.

[7] I. Gadaras and L. Mikhailov, "An interpretable fuzzy rule-based classification methodology for medical diagnosis," Artificial Intelligence in Medicine, vol. 47, no. 1, pp. 25–41, 2009.

[8] J. Abonyi and F. Szeifert, "Supervised fuzzy clustering for the identification of fuzzy classifiers," Pattern Recognition Letters, vol. 24, no. 14, pp. 2195–2207, 2003.

[9] M. Darzi, A. AsgharLiaei, M. Hosseini, and HabibollahAsghari, International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering, vol. 5, no. 5, pp. 220 – 223, 2011.

[10] V. N. Chunekar and H. P. Ambulgekar, "Approach of neural network to diagnose breast cancer on three different data set." in ARTCom. IEEE Computer Society, 2009, pp. 893–895.

[11] Q. Hu, J. Liu, and D. Yu, "Mixed feature selection based on granulation and approximation," Knowledge-Based Systems, vol. 21, no. 4, pp. 294–304, 2008.

[12] E. G¨urb¨uz and E. Kılıc¸, "A new adaptive support vector machine for diagnosis of diseases," Expert Systems, vol. 31, no. 5, pp. 389–397, 2014.

[13] H. Liu and H. Motoda, Feature Selection for Knowledge Discovery and Data Mining. Norwell, MA, USA: Kluwer Academic Publishers, 1998.

[14] A. Blum, A. Kalai, and J. Langford, "Beating the hold-out: Bounds for k-fold and progressive cross-validation," in Proceedings of the Twelfth Annual Conference on Computational Learning Theory, ser. COLT '99. New York, NY, USA: ACM, 1999, pp. 203–208.

[15] M. Lichman, "UCI machine learning repository," 2013.

[16] S. Aruna, S. Rajagopalan, and L. Nandakishore, "Knowledge based analysis of various statistical tools in detecting breast cancer," Computer Science & Information Technology, vol. 2, pp. 37–45, 2011.

[17] K. Wisaeng, "An empirical comparison of data mining techniques in medical databases," International Journal of Computer Applications, vol. 77, no. 7, pp. 23–27, September 2013, full text available.

[18] K. U. Rani, "Parallel approach for diagnosis of breast cancer using neural network technique," International Journal of Computer Applications, vol. 10, no. 3, pp. 1–5, 2010.

[19] L. Breiman and L. Breiman, "Bagging predictors," in Machine Learning, 1996, pp. 123–140.

[20] A. Graf, A. J. Smola, and S. Borer, "Classification in a normalized feature space using support vector machines," Neural Networks, IEEE Transactions on, vol. 14, no. 3, pp. 597–605, 2003.

[21] R. O. Duda and P. E. Hart, Pattern Classification and Scene Analysis. New Yotk: John Willey & Sons, 1973.

[22] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273–297, 1995.

[23] O. Chapelle and V. Vapnik, "Model selection for support vector machines," in Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29 – December 4, 1999], 1999, pp. 230–236.

[24] W. H. Wolberg, W. N. Street, D. M. Heisey, and O. L. Mangasarian, Computerized breast cancer diagnosis and prognosis from fine-needle spirates," Archives of Surgery, vol. 130, no. 5, pp. 511–516, 1995.

[25] T. Joachims, "Transductive inference for text classification using support vector machines," in Proceedings of the Sixteenth International Conference on Machine Learning, ser. ICML '99. San Francisco, USA.