# An Effective Intelligent Model for Medical Diagnosis

**Mohamed El-Rashidy [1], Taha Taha [2], Nabil Ayad [3] and Hoda Sroor [4]**

**[1] Dept. of Computer Science & Eng., Faculty of Electronic Engineering, Menoufiya University, Menouf, Egypt**

**[2] Dept. of Electronics& Electrical Communications, Faculty of Electronic Engineering, Menoufiya University, Menouf, Egypt**

**[3] Nuclear Research Center, Atomic Energy Authority, Cairo, Egypt.**

**[4] Dept. of Computer Science & Eng., Faculty of Electronic Engineering, Menoufiya University, Menouf, Egypt**

**Abstract**

A hybrid data mining model is proposed for finding an optimal number of different pathological types of any disease, and extracting the most significant features for each pathological type. This model is improved in order to reach the fewer subsets of features that have the most impact distinctive of each pathological type. This improvement is lead to the great importance in the decision making of the diagnosis process without confusion or ambiguity between the different variations of the diseases. This model and its optimization are based on fuzzy clustering, nearest neighbor classification, sequential backward search method, and averaging schema for features selection. Experiments have been conducted on three real medical datasets that have different diagnoses. The results show that the highest classification performance is obtained using our optimized model, and this is very promising compared to Naïvebayes, Linear and Polykernal Support Vector Machine (SVM), Artificial Neural Network (ANN), and Support Feature Machines (SFM) models.

***Keywords:*** *Data Mining, Fuzzy Clustering, Nearest Neighbor Classification, Features Selection.*

## 1. Introduction

Healthcare organizations are facing a major challenge in the patient diagnosis correctly and administering treatments that are effective. This challenge is related to the multiplicity of pathological types of diseases, which makes the diagnostic process more complex, especially if the symptoms and the results of the investigations indicated to these types are several and similarities. Therefore, it is important to find out an optimal number of different pathological types for each disease, and extracts the fewer subsets of features that have high classifiability for each type. This is due to the great importance of this information in the accuracy and speed of diagnostic process without confusion or ambiguity between the different variations of the diseases, and the need to avoid poor treatments that can lead to disastrous consequences. The practice of ignoring this vital knowledge leads to unwanted biases, errors and excessive medical costs which affect the quality of medical services that are provided to patients. This practice moved us toward the developing and optimizing data mining techniques, to get the most accurate knowledge that can be extracted from the medical databases to possess the highest quality of services. Data mining techniques have been successfully applied in various biomedical domains, for example the diagnosis and prognosis of cancers, liver diseases, diabetes, heart disease and other complex diseases [1-9]. These models are omitted the multiplicity of different pathological types of diseases in diagnosis process, and they deal with the disease as its one type and have only one set of distinctive features which distinguish it.

We proposed a hybrid approach based on fuzzy clustering, max-min, and feature selection models that employ extensive advances in classification medical data. We called this approach an Optimal Clustering for Support Feature Machine (OCSFM). The goal of OCSFM is to classify the disease into optimal number of classes, and select the fewer subsets of features that have high classifiability for each class. The advantage of OCSFM is that it uses fuzzy clustering that has classes with less sensitive to noise since noise data points will have very low degrees in all classes, which yields very accurate classification upon diagnosis.

OCSFM is tested on many diseases, similarities and multiplicities of features that are extracted for each of different pathological types of disease are founded, these practices may render the convergence impossible and are leading to random classification decisions. Therefore, we worked to derive an optimization for this model. This optimization is based on a new hybrid feature selection model that used averaging schema as a filter method, and sequential backward search as a wrapper method. The goal of this optimization is to extract the fewer subsets of features that have the most impact distinctive of each pathological type, and access the highest diagnostic accuracy in less time that provide the efficiency of treatment service.

We evaluated the performance of the optimized OCSFM model on the Wisconsin breast cancer (WBCD) [10], the Cleveland heart disease [10], and surgical patient's datasets compared to NaïveBayes [11], Linear SVM [12], Polykernal SVM [13], ANN [14], and SFM [9] models. The sections organization of this paper is as follows. In section 2, classification criteria's is described. Fuzzy clustering and averaging schema for feature selection are offered in section 3, and 4 respectively. In section 5, each step of OCSFM and our optimization is detailed. In section 6, the results and the performance characteristics of the proposed approach will be discussed. The concluding remarks are offered in section 7.

## 2. Classification Criteria's

The performance of data classification is commonly presented in terms of sensitivity and specificity. Sensitivity measures the fraction of positive test samples that are correctly classified as positive, then we define

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (1)$$

where TP and FN denote the number of true positives and false negatives, respectively. Specificity measures the fraction of negative test samples that are correctly classified as negative. Let FP and TN denotes the number of false positives and true negatives, respectively, then we define

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (2)$$

An overall accuracy is defined as

$$\text{Accuracy} = \frac{TN + TP}{TP + FP + TN + FN} \quad (3)$$

The Matthew's correlation coefficient (MCC) is a powerful accuracy evaluation criterion of machine learning methods. Especially, when the number of negative samples and positive samples are obviously unbalanced [1].

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4)$$

## 3. Fuzzy C-means Algorithm

Existing clustering models could be classified into three subcategories hierarchical, density based, and partition based approaches. Hierarchical algorithms organize objects into a hierarchy of nested clusters; hierarchical clustering can be divided into agglomerative and divisive methods [15-18]. Density based algorithms describe the density of data which are set by the density of its objects; the clustering involves the search for dense areas in the object space [19-21]. The idea of Partition based algorithms is to partition data directly into disjoint classes, this subcategory includes several algorithms as k-means, fuzzy c-means, P3M, SOM, graph theoretical approaches, and model based approaches [18] and [22-27]. These approaches assume a predefined number of classes. In addition, these approaches (except the fuzzy/possibilistic ones) always make brute force decisions on the class borders, for this, it may be easily biased by noisy data. This fact makes these fuzzy/possibilistic approaches less sensitive to noisy data.

Fuzzy c-means algorithm (FCM) is an iterative partitioning method [28]. It partitions data samples into c fuzzy classes, where each sample $x_j$ belongs to a class k with a degree of believe which is specified by a membership value $u_{kj}$ between zero and one such that the generalized least squared error function J is minimized.

$$J = \sum_{j=1}^{n} \sum_{k=1}^{c} (u_{kj})^m d(x_j, y_k) \quad (5)$$

Where m is a parameter of fuzziness, c is the number of classes, $y_k$ is the center of class k, and $d(x_j, y_k)$ expresses the similarity between the sample $x_j$ and the center $y_k$. The summation of the membership values for each sample is equal to one, and this guarantees that no class is empty.

$$0 < \sum_{k=1}^{c} u_{kj} \text{ And } \sum_{k=1}^{c} u_{kj} = 1 \quad \forall j = 1, \ldots, n \quad (6)$$

Because of calling this approach as a probabilistic clustering, since that the membership degrees for a given data point formally resemble the probabilities of its being a member of the corresponding class. This makes the possibilistic clustering less sensitive to noise since noise data points will have very low degrees in all classes. The

minimizations of J are resulted in the following membership function and class center.

$$u_{kj} = \frac{1}{\sum_{i=1}^{c}\left(\frac{d(x_j, y_k)}{d(x_j, y_i)}\right)^{\frac{2}{m-1}}} \quad (7)$$

where $u_{kj}$ is a possibility degree that measures how much typical is data point $x_j$ to class k. The membership degree of $x_j$ to a cluster not only depends on the distance between $x_j$ and that class, but also the distances between $x_j$ and the other classes. The partitioning property of a probabilistic clustering algorithm, which distributes the weight of $x_j$ on the different classes, is due to this equation. Although it is often desirable, the relative character of the membership degrees in a probabilistic clustering approach can lead to counterintuitive results.

$$y_k = \frac{\sum_{j=1}^{n}(u_{kj})^m x_j}{\sum_{j=1}^{n}(u_{kj})^m} \quad (8)$$

This choice makes $y_k$ proportional to the average intra class distance of k, and is related to the overall size and shape of the class.

## 4. Averaging Scheme

Feature selection algorithms could be classified into two subcategories; filter methods, and wrapper methods [29]. The filter methods estimate the classification performance by some indirect assessments such as distance measures which reflect how well the classes separate from each other. The wrapper methods based on a classifier to select the best subset of features that have the highest classification accuracy, these methods have many types in the searching process as sequential backward search (SBS), and sequential forward search (SFS).

Averaging scheme is a kind of filter methods. The selection feature of averaging scheme is based on two matrices. The first is an n×m intra class distance matrix D = ($d_{ij}$), and the other is an n×m inter class distance matrix $\overline{D} = (\bar{d}ij)$ . The entry of the intra class matrix $d_{ij}$ is the intra class distance, and the entry of the inter class matrix $\bar{d}ij$ is the inter class distance. After the two matrices are constructed, the selection of features is derived from the sum of intra class average distances ($d_{ij}$) are smaller than the sum of inter class average distances ($\bar{d}ij$) in the selected features [9].

## 5. OCSFM Model

The model is based on fuzzy C-means, max-min, and averaging schema to classify the data points into optimal number of representative classes, this representation is not aimed only to acquire less average distance (intra class distance), and highest average distance to all different classes (inter class distance), but it takes also into consideration the access to the highest classification accuracy. This model is an integration of both characteristics of supervised and unsupervised models, that makes OCSFM has classes less sensitive to noise, since it is of lowest noise data point's degree in all classes, and maximizes classification accuracy. The flowchart of OCSFM model is shown in Figure 1, where the inputs are the data set D = { $d_0, d_1, \ldots, d_n$ }, $c_{\min}$ and $c_{\max}$ are the minimal and the maximal numbers of expected clusters respectively.

In recapitulation, OCSFM model is composed of six main steps. The first step (Clustering), clusters data points in order to form optimal partitioning representation of classes with smallest intra class distance and greatest inter class distance using Fuzzy c-means algorithm. The second step (Selected Features), finds the optimal subset of features that have high classifiability for each class in order to have the maximum number of training samples correctly classified into those partitioning classes. The third step (Classification), training samples are classified according to those selected features by using nearest neighbor classifier, and computing the performance of data classification which is presented in terms of TP, TN, FP, and FN to obtain MCC. In the fourth step (Classes representatives points), FCM is sensitive to the initial center choices especially for noisy data. We use max-min approach [30], it is desirable to select the initial centers which are well separated; these centers make FCM classes separately groups in a feature space, it chooses a median of one class from those partitioning classes as a start point to select another classes representation points as separate as possible from start point. The fifth step (Multi step max-min algorithm), finds an optimal representative partitioning for a fixed number of classes, each iteration of the optimization process is based on clustering, selected features, and classification steps which is obtained by the max-min method but it changes start point with another class median. Iteration is stopped when each of classes medians is selected as a start point, therefore number of iteration for multi step max-min algorithm are equal to classes medians (number of classes). The sixth step

(Optimal classes number), computes an optimal classes number of partitioning classes by highest classification accuracy of the representation classes. For this, multi step max-min algorithm is repeated with increasing the number of partitioning classes from $c_{\min}$ to $c_{\max}$, using MCC as a validity measure in Equ.(4) which gives a better evaluation than overall accuracy with a lot of machine learning methods, such as SVM, ANN and BNN [1].

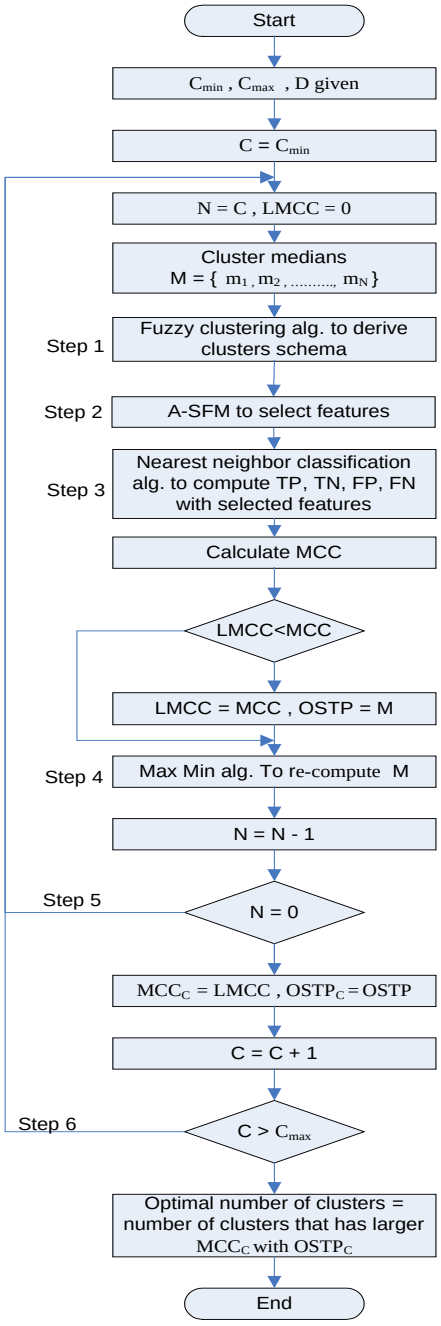


Fig. 1 Flowchart of OCSFM model.

### 5.1 Optimization of OCSFM Model

We improved OCSFM model in order to reach the highest classification accuracy in less time as possible, through access the fewer subsets of features that have the most impact distinctive of each class in the classification process. We proposed a new hybrid method for the feature selection instead of the averaging schema which used in the second step of the OCSFM model. This optimization is avoiding the irrelevant features that increase the computation time and may render the convergence between classes. This hybrid method consists of two steps which are explained in Figure 2. In the first step, averaging model is used as a filter method selecting the best subset of a given features set for each class, and reducing the number of features that have to be tested through the training of nearest neighbor classifier. In the second step, sequential backward search is used as a wrapper method; nearest neighbor classifier is used to select the subset of fewer features that has the highest classification accuracy for each class from the different subsets of features that are estimated from SBS method.

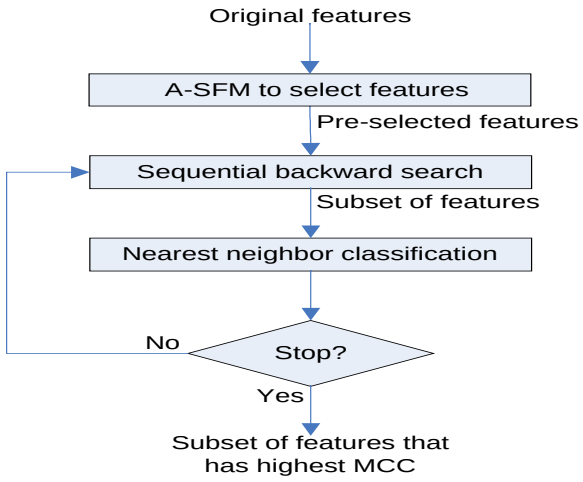


Fig. 2 The hybrid method of selected features.

## 6. Experimental Results and Discussion

All experiments were implemented and performed on AMD Phenom 9550 Quad Core 2.2 GHz workstation with 4 gigabytes of memory running on Windows Server 2003. The calculations and algorithms were implemented and run on ORACLE 10G. All programs were written by Java language. In the experiments, we apply our optimization model to diagnose several diseases that are related to breast cancer, heart disease, and post-operative infections. The first dataset acquired from the Breast Cancer Wisconsin Diagnostic (WDBC) database, they have been collected by Dr. William H. Wolberg at the University of

Wisconsin Madison Hospitals. There are 699 records in this database. Each record in the database has nine features which were computed from a digitized image of a fine needle aspirate of a breast mass. Those features, computed for each cell nucleus, are considered to be important characteristics for breast cancer diagnosis; those features include Clump thickness (Clump), uniformity of cell size (Ucellsize), uniformity of cell shape (Ucellshap), marginal adhesion (Mgadhesion), single epithelial cell size (Sepics), bare nuclei (Bnuclei), bland chromatin (Bchromatin), normal nucleoli (Normnuct), and Mitoses. In this database, 241 (34.5%) records are malignant and 458 (65.5%) records are benign.

The second dataset acquired from the Cleveland Heart Disease Database, they have been collected by Dr. Andras Janosi, at the Hungarian Institute of Cardiology. There are 297 records in this database; each record in the database has 13 features which are believed to be a good indicator for the angiographic disease status. Those features include chest pain type (Cp) (typical and a typical angina, non angina pain, and asymptomatic), resting blood pressure, serum cholesterol (Chol), resting electro cardio graphic results (Restecg) (normal, abnormality, probable), maximum heart rate (Thalach), indicator of exercise induced angina (Exang), Thal (normal, fixed detect, reversable detect), ST depression (Oldpeak), Slope of the peak exercise ST segment (Slope), number of major vessels colored by fluoroscopy (Ca), and the main criterion that physicians use to determine the diagnosis of heart disease is the narrowing in diameter of any major blood vessel. The diagnosis was considered to be positive (presence of heart disease) if the diameter of any major vessel was narrowed by more than 50%; and negative otherwise. In this database, 160 records (patients) have heart disease and 137 records (patients) have not heart disease.

The third dataset acquired from surgical patient's database, they have been collected from more than one server of Egyptian hospitals. There are 446 records in this database. Each record in the database has 15 features which are believed to be a good indicator for the infections. Those features include age, gender, clinical department name, operation name, operation risk index, health degree of patient (from 1 to 5), actual duration for operation, duration ideal for operation, wound class (none, mild, moderate, severe) of inflammation, length of stay sick before and after the operation, the period between first dose of anti biotic and starting operation, patient temperature during the operation, infection index (non-infected, infected), and name of organism that cause infection. In this database, 101 records (patients) have infection and 345 records (patients) have not infection.

We divided the data into training and testing phases, in test stage, 5-fold cross validation method was applied. First, we show the features that selected by OCSFM and optimization models, these features are summarized in Tables 1, 2 and 3. The tables shows the optimal number of different pathological types for each of the datasets, and the selected features for each of different pathological types of the disease by using both of averaging schema and our hybrid model, where the number of negative and positive classes reflects the number of different pathological types injured and not injured by a disease on respectively. Multiplicity of negative classes is points to existence of many different subsets of features that are caused the disease, but the diversity of suspected cases reflects the multiplicity of positive classes. The results show that similarities of the subsets of multiplicity features in the different pathological types of a disease by using OCSFM model, which leading to the confusion or the ambiguity in the diagnostic process. The distinctive subsets of the fewer features for each disease type are extracted by using our optimization model, which lead to the altitude of diagnostic accuracy in less time as possible.

Second, the altitude of diagnostic accuracy using optimization model can be appeared by sensitive rates comparing with NaïveBayes, Linear SVM, Polykernal SVM, artificial ANN, SFM, OCSFM models in Tables 4, 5 and 6. The results show that our optimization model achieves the highest classification accuracy, which leads to provide the efficiency of treatment service, helps the pathologist to better detect the type of tumor (benign or malignant), the avoidance of diseases complication, chemotherapy complication, exposure to radiation, and mastectomy. And also the improvement is considered as the important purpose that can help physicians to better detect heart disease and post-operative infections.

Table 1: Selected features for each of different pathological types of Breast Cancer Dataset used OCSFM and optimized OCSFM models.

| ***Optimal classes number*** | ***Negative classes number*** | ***Positive classes number*** |
|---|---|---|
| 14 | 10 | 4 |
| The main features for each negative class | | |

| ***Class No.*** | ***OCSFM*** | ***No. of Feat.*** | ***Optimized OCSFM*** | ***No. of Feat.*** |
|---|---|---|---|---|
| 1 | Clump, Ucellsize, Ucellshap, Mgadhesion, Sepics, Bnuclei, Bchromatin, and Normnuct. | 8 | Ucellsize, Ucellshap, Bchromatin, Bnuclei, Normnuct. | 5 |
| 2 | Clump, Ucellsize, Ucellshap, Mgadhesion, Sepics, | 9 | Clump, Ucellsize, Ucellshap, Sepics, Mitoses. | 5 |

| | | | | |
|---|---|---|---|---|
| | Bnuclei, Bchromatin, Normnuct, and Mitoses. | | | |
| 3 | Clump, Ucellsize, Ucellshap, Mgadhesion, Sepics, Bnuclei, Bchromatin, Normnuct, and Mitoses. | 9 | Ucellsize, Ucellshap, Sepics, Bnuclei, Bchromatin, Normnuct, Mitoses. | 7 |
| 4 | Clump, Ucellsize, Ucellshap, Mgadhesion, Sepics, Bnuclei, Bchromatin, and Normnuct. | 8 | Bnuclei. | 1 |
| 5 | Clump, Ucellsize, Ucellshap, Mgadhesion, Sepics, Bnuclei, Bchromatin, and Normnuct. | 8 | Ucellsize, Ucellshap, Sepics, Bnuclei, Bchromatin, Normnuct. | 6 |
| 6 | Clump, Ucellsize, Ucellshap, Sepics, Bnuclei, Bchromatin, Normnuct, and Mitoses. | 8 | Ucellsize, Ucellshap, Sepics, Bnuclei, Bchromatin, Normnuct, Mitoses. | 7 |
| 7 | Clump, Ucellsize, Ucellshap, Mgadhesion, Sepics, Bnuclei, Bchromatin, and Normnuct. | 8 | Ucellsize, Sepics, Normnuct. | 3 |
| 8 | Clump, Ucellsize, Ucellshap, Bnuclei, Bchromatin, and Normnuct. | 6 | Bnuclei, Bchromatin, Normnuct. | 3 |
| 9 | Clump, Ucellsize, Ucellshap, Mgadhesion, Sepics, Bnuclei, and Bchromatin. | 7 | Ucellsize, Sepics, Bnuclei, Bchromatin. | 4 |
| 10 | Clump, Ucellsize, Ucellshap, Mgadhesion, Sepics, Bnuclei, Bchromatin, and Normnuct. | 8 | Ucellsize, Ucellshap, Sepics, Bchromatin, Normnuct. | 5 |

Table 2: Selected features for each of different pathological types of Heart Disease Dataset used OCSFM and optimized OCSFM models.

| ***Optimal classes number*** | ***Negative classes number*** | ***Positive classes number*** |
|---|---|---|
| 3 | 1 | 2 |
| The main features for each negative class | | |

| ***Class No.*** | ***OCSFM*** | ***No. of Feat.*** | ***Optimized OCSFM*** | ***No. of Feat.*** |
|---|---|---|---|---|
| 1 | Cp, Oldpeak, Ca, Slope, Thalach, Thal, Exang, Restecg, Chol. | 9 | Ca, Slope, Thal, Exang. | 4 |

Table 3: Selected features for each of different pathological types of Surgical Patient's Dataset used OCSFM and optimized OCSFM models.

| ***Optimal classes number*** | ***Negative classes number*** | ***Positive classes number*** |
|---|---|---|
| 5 | 2 | 3 |
| The main features for each negative class | | |

| ***Class No.*** | **OCSFM** | **No. of Feat.** | **Optimized OCSFM** | **No. of Feat.** |
|---|---|---|---|---|
| 1 | age, gender, clinical department name, operation name, operation risk index, health degree of patient, actual duration for operation, duration ideal for operation, wound class of inflammation, length of stay sick before and after the operation, the period between first dose of anti biotic and starting operation, patient temperature during the operation | 13 | operation risk index. | 1 |
| 2 | age, gender, clinical department name, operation name, operation risk index, health degree of patient, actual duration for operation, duration ideal for operation, wound class of inflammation, length of stay sick before and after the operation, the period between first dose of anti biotic and starting operation, patient temperature during the operation | 13 | duration ideal for operation, actual duration for operation, operation risk index | 3 |

Table 4: Training and testing performance in % sensitivity, specificity, overall accuracy and MCC of NaïveBayes, Linear SVM, Polykernal SVM, ANN, SFM, OCSFM and Optimized OCSFM approaches for Diagnosis of Breast Cancer in WDBC Database.

| ***Classification algorithm*** | Training Data | | | | Testing Data | | | |
|---|---|---|---|---|---|---|---|---|
| | Sens. | Spec. | Accu. | MCC | Sens. | Spec. | Accu. | MCC |
| NaïveBayes | 97.19 | 97.09 | 97.12 | 94.05 | 95.55 | 97.64 | 96.92 | 93.20 |
| Linear SVM | 94.54 | 94.60 | 94.56 | 88.17 | 94.73 | 22.22 | 80.85 | 23.91 |
| Polykernal SVM | 97.59 | 96.26 | 97.14 | 93.68 | 93.33 | 97.64 | 96.15 | 91.47 |
| ANN | 97.37 | 98.75 | 97.85 | 95.33 | 97.77 | 97.64 | 97.69 | 94.94 |
| SFM | 97.85 | 91.32 | 95.60 | 90.22 | 97.64 | 88.88 | 94.61 | 88.03 |
| OCSFM | 97.31 | 98.97 | 97.89 | 95.42 | 97.64 | 97.77 | 97.69 | 94.94 |
| Optimized OCSFM | 97.58 | 98.48 | 98.24 | 96.19 | 96.47 | 100 | 97.71 | 95.10 |

Table 5: Training and Testing Performance in % sensitivity, specificity, overall accuracy and MCC of NaïveBayes, Linear SVM, Polykernal SVM, ANN, SFM, OCSFM and Optimized OCSFM approaches for Diagnosis of Cleveland Heart Disease Database.

| ***Classification algorithm*** | Training Data | | | | Testing Data | | | |
|---|---|---|---|---|---|---|---|---|
| | Sens. | Spec. | Accu. | MCC | Sens. | Spec. | Accu. | MCC |
| NaïveBayes | 79.56 | 88.12 | 84.17 | 68.14 | 72.72 | 74.28 | 73.68 | 46.12 |
| Linear SVM | 80.29 | 90.00 | 85.52 | 70.89 | 77.27 | 88.57 | 87.71 | 67.99 |
| Polykernal SVM | 79.56 | 89.37 | 84.84 | 69.53 | 80.36 | 85.71 | 82.96 | 69.07 |
| ANN | 86.86 | 88.75 | 87.87 | 75.61 | 77.27 | 88.57 | 84.21 | 66.45 |
| SFM | 82.60 | 85.60 | 84.16 | 68.26 | 74.28 | 95.45 | 82.45 | 67.99 |
| OCSFM | 86.08 | 88.80 | 87.50 | 74.94 | 82.85 | 90.90 | 85.96 | 72.10 |
| Optimized OCSFM | 88.69 | 89.60 | 89.16 | 78.29 | 84.53 | 93.71 | 87.72 | 76.35 |

Table 6: Training and Testing Performance in % sensitivity, specificity, overall accuracy and MCC of NaïveBayes, Linear SVM, Polykernal SVM, ANN, SFM, OCSFM and Optimized OCSFM approaches for Diagnosis of Surgical Patient's Database.

| ***Classification algorithm*** | Training Data | | | | Testing Data | | | |
|---|---|---|---|---|---|---|---|---|
| | Sens. | Spec. | Accu. | MCC | Sens. | Spec. | Accu. | MCC |
| NaïveBayes | 92.72 | 72.13 | 89.91 | 60.57 | 92.11 | 55.55 | 85.11 | 49.90 |
| Linear SVM | 99.74 | 26.22 | 89.68 | 46.60 | 100 | 11.11 | 82.97 | 30.29 |
| Polykernal SVM | 99.74 | 24.59 | 89.46 | 44.95 | 100 | 11.11 | 82.97 | 30.29 |
| ANN | 100 | 39.34 | 91.70 | 59.91 | 100 | 22.22 | 85.11 | 43.32 |
| SFM | 90.75 | 39.62 | 83.95 | 30.37 | 94.74 | 11.11 | 78.72 | 9.41 |
| OCSFM | 95.08 | 71.69 | 91.97 | 65.74 | 94.74 | 66.66 | 89.36 | 64.29 |
| Optimized OCSFM | 98.26 | 89.28 | 97.01 | 87.55 | 97.22 | 93.89 | 95.74 | 86.25 |

## 7. Conclusions

In this paper, the OCSFM approach has been improved and applied to the tasks of breast cancer, heart diseases, and post-operative infections diagnosis. Results are indicated that our proposed approach found out the optimal number of different pathological types of these diseases, and extracted the fewer subsets of features that have the most impact distinctive of each pathological type without confusion or ambiguity between the different variations of these diseases. Here, after applying the proposed approach, the accuracy of the diseases diagnosis in WDBC, Cleveland Heart Disease, and surgical patient's datasets has been improved by sensitive rates.

## References

[1] H. Cheng, J. Shan, W. Ju, Y. Guo, and L. Zhang, "Automated breast cancer detection and classification using ultra sound images: Asurvey", Pattern Recognition, 43, 299-317, 2010.

[2] B. Riccardo, and Blaz Z., "Predictive data mining in clinical medicine: Current issues and guidelines", international journal of medical informatics, 77, 81-97, 2008.

[3] L. Rong-Ho, "An intelligent model for liver disease diagnosis", Artificial Intelligence in Medicine, 47, 53-62, 2009.

[4] H. Yue, M. Paul, B. Norman, and H. Roy, "Feature selection and classification model construction on type 2 diabetic patient's data", Artificial Intelligence in Medicine, 41, 251-262, 2007.

[5] M. Choua, T. Leeb, Y. Shaoc, and I. Chenb, "Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines", Expert Systems with Applications, 27, 133-142, 2004.

[6] J. Elmore, M. Wells, M. Carol, H. Lee, D. Howard, and A. Feinstein, "Variability in radiologists interpretation of mammograms", New England Journal of Medicine, 331(22), 1493-1499, 1994.

[7] A. Mehmet, "Support vector machines combined with feature selection for breast cancer diagnosis", Expert Systems with Applications, 36, 3240-3247, 2009.

[8] M. Ilias, Z. Elias, and A. Ioannis, "An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers", Appl Intell, 30, 24-36, 2009.

[9] Ya-Ju F., and Wanpracha A. Ch., "Optimizing feature selection to improve medical diagnosis", Ann Oper Res, 174, 169-183, 2010.

[10] Cleveland Heart Disease and Wisconsin Breast Cancer Datasets are originally available on UCI Machine Learning Repository website http://archive.ics.uci.edu.

[11] P. Bhargavi, and S. Jyothi, "Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils", International Journal of Computer Science and Network Security, 9(8), 117-122, 2009.

[12] L. Zhizheng, and Z. Tuo, "Feature selection for linear support vector machines", The 18th International Conference on Pattern Recognition IEEE, 2006.

[13] I. Bhattacharya, and M. P. S. Bhatia, "SVM classification to distinguish Parkinson disease patients", A2CWiC '10 Amrita ACM-W Celebration on Women in Computing in India, 2010.

[14] J. Paulo, and F. Azzam, "The use of artificial neural networks in decision support in cancer: A systematic review ", Neural Networks, 19( 4), 408-415, 2006.

[15] M. Eisen, P. Spellman, P. Brown, and D. B otstein, "Cluster analysis and display of genome wide expression patterns", Natl Acad Sci USA, 95(25), 14863-14868, 1998.

[16] M. Blatt, S. Wiseman, and E. Domany, "Super-paramagnetic clustering of data", Phys Rev Lett, 76, 3251-3254, 1996.

[17] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems", IEEE, 86(11), 2210-2239, 1998.

[18] J. Herrero, A. Valencia, and J. Dopazo, "A hierarchical unsupervised growing neural network for clustering gene expression patterns", Bioinformatics, 17(2), 126-136, 2001.

[19] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: a survey", IEEE Trans Knowl Data Eng, 16(11), 1370-1386, 2004.

[20] D. Jiang, J. Pei, and A. Zhang, "DHC: a density-based hierarchical clustering method for time series gene expression data", the 3rd IEEE symp on bioinformatics and bioengineering, Maryland, USA, 393-400, 10-12 March 2003.

[21] A. Hinneburg, and D. Keim, "An efficient approach to clustering in large multimedia database with noise", the 4th int conf on knowledge discovery and data mining, NY, USA, 58–65, 27–31 August 1998.

[22] W. Au, K. Chan, A. Wong, and Y. Wang, "Attribute clustering for grouping, selection, and classification of gene expression data", IEEE/ACMTrans Comput Biol Bioinform, 2(2), 83–101, 2005.

[23] D. Bickel, "Robust cluster analysis of microarray gene expression data with the number of clusters determined biologically", Bioinformatics, 19(7), 818–824, 2003.

[24] R. Guthke, W. Schmidt-Heck, D. Hann, and M. Pfaff, "Gene expression data mining for functional genomics", the European symp on intel techn, Aachen, Germany, 170–177, 2000.

[25] L.B. Romdhane, H. Shili, and B. Ayeb," Mining microarray gene expression data with unsupervised possibilistic clustering and proximity graphs", Appl Intell, 10.1007/s, 10489-009, 2009.

[26] R. Shamir, and R. Sharan, "CLICK: A clustering algorithm for gene expression analysis", the int conf on intelligent systems for molecular biology, CA, USA, 307–316, 19–23 August 2000.

[27] K. Yeung, C. Fraley, A. Murua, A. Raftery, and W. Ruzz, "Model-based clustering and data transformations for gene expression data", Bioinformatics, 17(10), 977–987, 2001.

[28] J. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms", New York: Plenum, 1981.

[29] L. Ming-Chi, "Using support vector machine with a hybrid feature selection method to the stock trend prediction", Expert Systems with Applications, 36, 10896–10904, 2009.

[30] J. Tou, and R. Gonzalez, "Pattern recognition principles", Addison-Wesley, Reading, 1974.

**First Author** obtained his Master degree in computer science and engineering, 2008. Currently, he is working as a Lecturer Assistant in the Dept. of Computer Science and Engineering, Faculty of Electronic Engineering, 32952, Menouf, Menoufiya University -Egypt. Areas of interest of the author include data mining and bioinformatics.

**Second Author** was born in Tanta, Egypt, on October 11, 1946. He received the B.Sc. degree (with distinction) in communication engineering from Menoufiya University, Egypt, in June 1969, the M.Sc. degree in communication engineering from Helwan University, Egypt, in April 1978, and the Ph.D. degree (very honorable) in electronic engineering from the National Polytechnic Institute, Toulouse, France, in June 1985. From September 1969 to July 1978, he was a Demonstrator, in July 1978, he was an Assistant Lecturer, in November 1985, he was a Lecturer, in February 1990, he was an Assistant Professor, and in September 1995, he was named Professor, all in the Faculty of Electronic Engineering, Menoufiya University, Communication Department,. He was appointed Vice Dean from February 2002 to October 2005, and Head of the Communication Department, from November 2005 to July 2007. At present, he is an Emeritus Professor at the same department. His main research interests are surface acoustic wave devices, optical devices, superconductor devices, medical applications of ultrasound, and bioinformatics.

**Third Author** received Ph.D degree in CSE from Cairo University, in 1984. He is working as vice chairman for reactors division, Nuclear Research Center, Atomic Energy Authority- Egypt. He is a member of IEEE. His main research interests database and networks.

**Fourth Author** received Ph.D degree in CSE from Menoufiya University, in 1991. She is working as Professor in Dept. of Computer Science and Engineering, Faculty of Electronic Engineering, 32952, Menouf, Menoufiya University- Egypt, her main research interests parallel processing and database.