An Effective Intelligent Model for Finding an Optimal Number of Different Pathological Types of Diseases

Mohamed A. El-Rashidy
Dept. of Computer Science & Eng.,
Faculty of Electronic Engineering,
Menoufiya University, Egypt.

Taha E. Taha
Dept. of Electronics& Electrical Communications,
Faculty of Electronic Engineering,
Menoufiya University, Egypt.

Nabil M. Ayad
Nuclear Research Center,
Atomic Energy Authority,
Cairo, Egypt.

ABSTRACT
A new hybrid data mining model is proposed to provide a comprehensive analytic method for finding an optimal number of different pathological types of any disease and its complications, an optimal partitioning representative and extracts the most significant features for each pathological type. This model is an integration of both characteristics of supervised and unsupervised models and is based on clustering, feature selection, and classification concepts. This model takes into consideration access to the highest classification accuracy during the clustering process. Experiments have been conducted on 3 real medical datasets related to the diagnosis of breast cancer, heart disease, and post-operative infections. The performance of this method is evaluated using information entropy, squared error, classification sensitivity, specificity, overall accuracy, and Matthew's correlation coefficient. The results show that the highest classification performance is obtained using our proposed model, and this is very promising compared to NaïveBayes, Linear Support Vector Machine (Linear SVM), Polykernal Support Vector Machine (Polykernal SVM), Artificial Neural Network (ANN), and Support Feature Machines (SFM) models.

General Terms
Data Mining, Diagnostic and Decision Support System.

Keywords
Clustering, feature selection, classification, SFM model, breast cancer, heart disease, post-operative infection.

1. INTRODUCTION
Each disease has a number of different pathological types and has distinguished features for each of them. These features include symptoms and results of the investigations required to indicate the disease. This information is usually taken into consideration by the physician for the diagnosis process to determine the appropriate method of treatment. There are many methods of treatment for each disease and the choice of the specific method depends on the type of pathology and the extent of its complications. Therefore, it is important to know the optimal number of the different pathological types and the complications of each disease, and the impact of these significant features for each type. This is due to the great importance of this information in the diagnostic accuracy 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 service that are provided to patients.

Medical databases include rich data that are the basis of useful knowledge. Analyzing and mining these databases for clinical decision support is a task of great importance to minimize the risk of making wrong decisions in diagnosis and treatment. The goal of predictive data mining in clinical medicine is to derive models that can use patient's specific information to predict the outcome of interest that supports clinical decision making. Data mining techniques have been successfully applied to various biomedical domains, for example the diagnosis and prognosis of cancers, liver diseases, diabetes, heart diseases and other complex diseases [1-4]. Classification, clustering, and feature selection are important data mining techniques widely used in numerous real world applications. Classification is the process of finding a model that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data whose class label is known. Classification models have a property of supervised learning, which analyze class labeled data objects. Clustering analyzes data objects without consulting a known class label. It can be used to generate such labels, so it has unsupervised learning properties. The objects are clustered based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Each cluster that is formed can be viewed as a class of objects, from which rules can be derived. Feature selection is used to select the best subset of features for each class which maximizes the classification accuracy despite the use of impact effective features to make the decision.

There are many techniques applied in medical diagnoses, such as artificial neural network, multivariate adaptive regression splines, decision trees, support vector machines, Bayesian, Support Feature Machine [5-9]. These models are omitted in the diagnosis process of different pathological types of disease, and they deal with the disease as one type and have only one set of distinctive features distinguishing it. In this paper we present a new hybrid approach based on fuzzy clustering, max-min, and SFM models that employ advances in classification of medical data. We call this hybrid approach an Optimal Clustering for Support Feature Machine (OCSFM). The goal of OCSFM is to classify the disease into optimal number of classes (different pathological types of disease and its complications), optimal

representative of these classes, and select the subset of features that have high classifiability for each class, which reflect the diversity of the disease types. 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.

We evaluated the performance of OCSFM 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 organization of the sections in this paper is as follows. In section 2, results of the clustering methods' survey are offered. In section 3, SFM and classification criteria will be described. In section 4, each step of our proposed OCSFM will be detailed. In section 5, the results and performance characteristics of the proposed approach will be discussed. The concluding remarks will be offered in section 6.

2. FUZZY CLUSTERING

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.

2.1 Fuzzy C-means Algorithm

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 belief which is specified by a membership value u_{kj} between zero and one such that the generalized least squared error function J is minimized.

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

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.

$$0 < \sum_{k=1}^{c} u_{kj} \text{ And } \sum_{k=1}^{c} u_{kj} = 1 \;\; \forall j = 1,....,n \qquad (2)$$

This approach is a probabilistic clustering, since 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 the following membership function and class center:

$$u_{kj} = \frac{1}{\sum_{i=1}^{c}\left(\frac{d(x_j, y_k)}{d(x_j, y_i)}\right)^{\frac{2}{m-1}}} \qquad (3)$$

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 that the relative character of the membership degrees in a probabilistic clustering approach can lead to counterintuitive results.

$$y_k = \frac{\sum_{j=1}^{n}(u_{kj})^{m} x_j}{\sum_{j=1}^{n}(u_{kj})^{m}} \qquad (4)$$

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.

3. SUPPORT FEATURE MACHINE ALGORITHM

Support feature machine is a classification method which uses the nearest neighbor rule to integrate spatial and temporal properties, and formulates an optimized model to select a group of features $m_s \le m$ that give the best discrimination under the nearest neighbor rule which maximizes the number of correctly classified samples or minimizes the classification error [9]. Nearest neighbor rule has two improved schema on unbalanced data, voting under distances measure (voting schema) and directly comparing averaged distances (averaging schema). These schemas have the same class samples which are close to each other and are away from the different class as much as possible with the selected features.

3.1 Voting Scheme (V-SFM)

The selection feature of the voting scheme is based on one matrix A = (a_{ij}) with an n × m, i = 1,…, n and j = 1,…, m, where n is the number of samples and m is the number of features. The classification is correct when the average distances from sample i to all other samples in the same class at feature j (intra-class distance) is smaller than the average distances to all samples in different classes at the same feature (inter-class distance). Therefore, the entry a_{ij} = 1 indicates that the nearest neighbor rule is a correctly classified sample i at feature j, otherwise a_{ij} =0. The best subset of features is selected, which gives the majority correct votes (value 1's) that have the maximum number of correct classified samples [9].

3.2 Averaging Scheme (A-SFM)

The selection feature of the 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 $\overline{D} = (\overline{d}ij)$. The entry of the intra-class matrix dij is the intra-class distance, and the entry of the inter-class matrix $\overline{d}ij$ is the inter-class distance. After the two matrix 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 ($\overline{d}ij$) in the selected features [9].

3.3 Classification Criterion

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

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (5)$$

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 denote the number of false positives and true negatives, respectively. Then, we define

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (6)$$

An overall accuracy is defined as

$$\text{Accuracy} = \frac{TN + TP}{TP + FP + TN + FN} \quad (7)$$

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].

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (8)$$

4. PROPOSED MODEL

We propose a new hybrid algorithm based on fuzzy C-means, max-min, and support feature machine to classify the data points into optimal number of representative classes, this representative 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 into consideration access to the highest classification accuracy. This model integrates the characteristics of both supervised and unsupervised models, which makes the OCSFM has classes less sensitive to noise, and maximizes the classification accuracy. The flowchart of our proposed model is given in Fig. 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 maximal numbers of expected clusters, respectively.

Our model is composed of six main steps. In the first step (Clustering), data points are clustered in order to form optimal partitioning representative of classes with smallest intra-class distance and greatest inter-class distance using Fuzzy c-means algorithm. In the second step (Selected Features), the optimal subset of features that have high classifiability for each class is found in order to have the maximum number of training samples correctly classified into those partitioning classes.

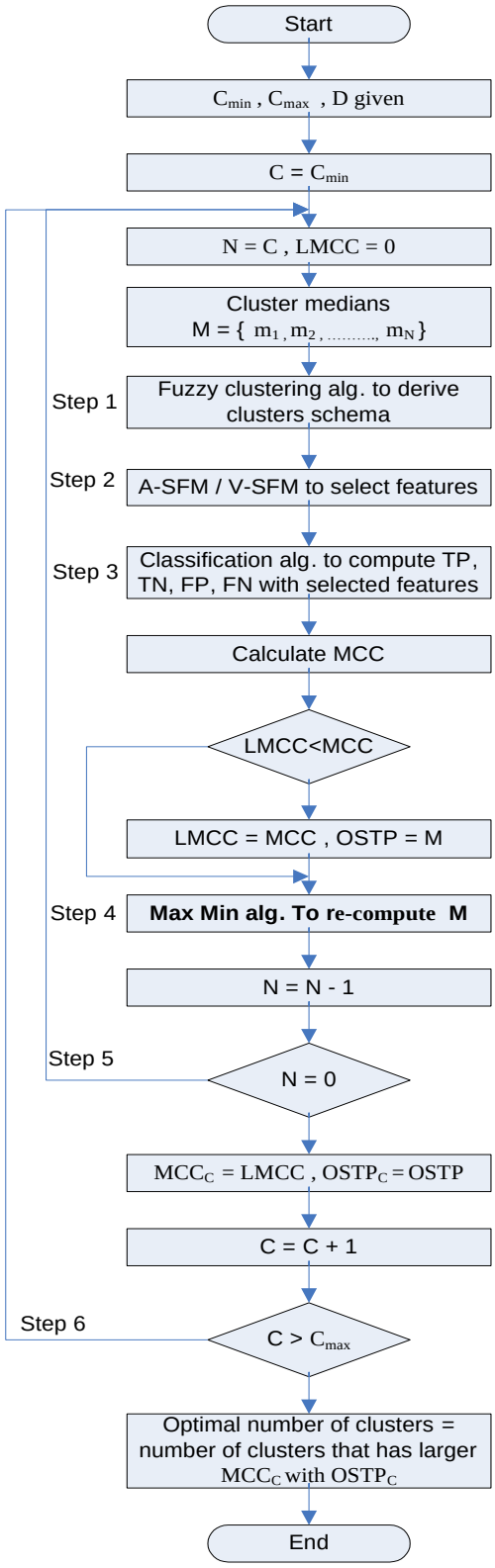

Fig 1: Flowchart of our proposed algorithm (OCSFM)

In the third step (Classification) training samples are classified according to those selected features, and the performance of data classification is computed and presented in terms of TP, TN, FP,

and FN to obtain MCC. In the fourth step (Classes representative points), Fuzzy c-means algorithms (FCM) are sensitive to the initial center choices especially for noisy data. We use max-min approach [29], where it is desirable to select the initial centers that are well separated. These centers make FCM classes as separate groups in a feature space, it then chooses a median of one class from those partitioning classes as a start point to select another classes representative points as separate as possible from the start point. In the fifth step (Multi step max-min algorithm), find 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 are obtained by the max-min method but it changes the start point with another class median. Iteration is stopped when each of the class medians has been selected as a start point, therefore the number of iterations for multi-step max-min algorithm is equal to the class medians (number of classes). In the sixth step (Optimal classes number), compute an optimal number of partitioning classes by highest classification accuracy of the representative 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. (8), which gives a better evaluation than overall accuracy with a lot of machine learning methods, such as SVM, ANN and BNN [1].

4.1 OCSFM Algorithm

In our proposed approach (OCSFM) in Algorithm 1 for a given number of classes $c(c_{\min} \le c \le c_{\max})$, We find the optimal partitioning c_o using the modified multi step max-min algorithm, and use MCC as a validity measure in Equ.(8), the optimal number of classes produces the highest MCC.

Algorithm 1: OCSFM algorithm

Input: Data Set D = { $d_0, d_1, \ldots, d_n$ }, $c_{\min}$ and $c_{\max}$ the minimal and maximal numbers of expected clusters, respectively.

Output: c_o An optimal cluster scheme.

begin

Let C ← $c_{\min}$.

while (c <= $c_{\max}$) **do**

1. Randomly choose a data point m from D as the starting point.
2. Perform Max-Min algorithm (Algorithm 3) to compute the cluster medians M={ $m_1, \ldots, m_c$ }.
3. Perform Multi step Max-Min algorithm (Algorithm 2) to find the optimal partitioning for C clusters with highest MCC.
4. Let C = C + 1.

end while

Let c_o ← the optimal partitioning with the highest MCC.

return (c_o)

end

4.2 Multistep Max-Min Algorithm

The multistep max-min algorithm is used to find an optimal representation of partitioning classes for a fixed number of classes. In the multistep max-min algorithm, each iteration of the process is based on the partitioning obtained by the max-min method. The max-min method tries to select class representatives by making classes as separate as possible. The basic max-min approach was first proposed in [29] and is summarized in Algorithm 3. Our approach is summarized in Algorithm 2.

Initially, we compute a class scheme using the max-min method starting from the given initial point m. Thereafter, a refinement of this scheme is performed using the same max-min method but with the computed object medians as new starting points. c_m in c_i is called an object median of c_i if:

$$d(c_m, m_i) = \min_{x \in c_i} d(x, m_i) \qquad (9)$$

Where m_i is the mean of class c_i .

Algorithm 2: Multi step Max-Min approach with MCC as an optimization criterion

Input: Data set D = { $d_0, d_1, \ldots, d_n$ }, number of clusters C, starting point m .

Output: Optimal partitioning c_o = {C1,…, Cc} for a fixed number of classes.

begin

1. Compute the cluster representative for c_o using Algorithm 4
2. **for** $i \leftarrow 2$ **to** c **do**

3.1.Recompute Cluster medians with m_i as a start point using Algorithm 3.

3.2. Recompute the cluster representative for $\bar{c}_o$ using Algorithm 4

3.3. Perform selected features algorithm (Algorithm 5 for V-SFM or Algorithm 6 for A-SFM) to find optimal separability features F ={ $f_1, \ldots, f_n$ }for c_o and $\bar{c}_o$.

3.4. Perform nearest neighbor classification algorithm (Algorithm 7) to find TP, TN, FP, and FN for c_o and $\bar{c}_o$.

3.5. Compute the MCC of both cluster schemes c_o and $\bar{c}_o$ using Equ. (9).

3.6. **if** (MCC($\bar{c}_o$) > MCC(c_o)) **then** Let $c_o \leftarrow \bar{c}_o$

end for

end

Algorithm 3: Max-Min approach

Input: Data set D = { $d_0, d_1, \ldots, d_n$ }, number of clusters C, starting point m .

Output: Cluster medians M = { $m_1, \ldots, m_c$ }.

begin

1. Let $m_1 \leftarrow m$, and M = { m_1 }.
2. **for** $i \leftarrow 2$ **to** c **do**

2.1. **for all** $\forall d_h \in D \setminus M$ **do**
Compute all distances $dis(d_h, m_j) \forall m_j \in M$ Save only the minimum distance in a set DIS
end for
2.2. Compute m_i the data point with the maximum distance value in DIS
2.3. Let $M \leftarrow M \cup \{m_i\}$
end for
return (M)
end

4.3 Representatives of Classes

After obtaining the new set of representatives using Algorithm 2, in Algorithm 4 each data point is assigned to the cluster that has the nearest representative (median). The whole process is repeated until the set of representatives becomes stable, or a maximal number of iterations are reached.

Algorithm 4: Classes representatives approach
Input: Data set D = { $d_0, d_1, \ldots, d_n$ }, Clusters medians M = { $m_1, \ldots, m_c$ }.
Output: New Cluster medians $\overline{M}$ = { $\overline{m}_1, \ldots, \overline{m}_c$ }.
begin
flag ← off.
while flag = off **do**
1. **for each** $d_j \in D$ **do**
choose the nearest representative, say m_i group d_j into cluster c_i (whose representative is m_i).
2. **for** i ← 1 **to** c **do**
compute $\overline{m}_i$, the object median for $\overline{c}_i$ as its new representative using Equ. (9).
3. Let $\overline{M}$ = { $\overline{m}_1, \ldots, \overline{m}_c$ }.
4. **if** (M = $\overline{M}$) OR (maximal iterations are reached) **then**
flag ← on.
else flag ← off.
end if
end while
end

4.4 Selected Features

The classification accuracy of every feature can be evaluated by applying A-SFM or V-SFM that gives the accuracy information for all features. These two algorithms are formulated to incorporate all the classification decisions made by each feature and select the best subset of features which maximizes the classification accuracy. We supported our proposed approach with using the A-SFM model in Algorithm 5. It is seen as having a better performance than the V-SFM [9].

Algorithm 5: Selected features A-SFM approach
Input: Optimal partitioning c_o = {C1,…, Cc} that have j = 1,…, m features, positive and negative point declaration.
Output: optimal separability features F = { $f_1, \ldots, f_n$ }
begin
1. Let POS set is empty.
2. **for** i ← 1 **to** c **do**
2.1. **if(** number of positive data points for c_i > number of negative data points for c_i) **then** POS = POS $\cup$ C_i .
end if
end for
3. **for** i ← 1 **to** c **do**
3.1. **if(** number of negative data points for c_i > number of positive data points for c_i) **then**
begin
3.1.1. Compute intra-class matrix D = (dij) (average distance between each data point i, in c_i , and all other data points in the same class for each feature j).
3.1.2 .Compute inter-class matrix $\overline{D} = (\overline{dij})$ (average distance between each data point i, in c_i , and all data points in POS set for each feature j).
end if
end for
4. **for** i ← 1 **to** c **do**
4.1. **if(** number of negative data points for C_i > number of positive data points for C_i) **then**
begin
4.1.1. f_i ← empty.
4.1.2. **for** j ← 1 **to** m **do**
4.1.2.1 **if (** $\sum d_{ij} \lhd \sum \overline{d}_{ij}$) **then** $f_i = f_i \cup$ j.
end if
end for
4.1.3 F=F $\cup$ f_i .
end if
end for
return(F)
end

4.5 Classification

After obtaining the optimally selected features, the dataset is classified according to the selected features F using Algorithm 6. V-SFM classifies an unlabeled class sample to the class with majority voting from all selected features. A-SFM classifies each of the data points to the class whose baseline training samples are more similar to it, based on the selected features. After each data point is labeled by SFM schemes, accuracies of SFM schemes can be calculated by comparing the labeled class with the actual class of each sample [9].

Algorithm 6: Nearest neighbor classification approach
Input: Data set D = { $d_0, d_1, \ldots, d_n$ }, Optimal partitioning c_o = {C1,…, Cc}, optimal separability features F = { $f_1, \ldots, f_n$ }.

Output: TP, TN, FP, and FN the true positive, true negative, false positive, and false negative, respectively.
begin
1. **for each** $d_j \in D$ **do**
begin
1.1. choose the nearest neighbor according to features F for each class, say C_i .
1.2. **if(** number of negative data points for C_i > number of positive data points for C_i) **then**
if(d_j is a negative data point) **then** TN=TN+1
else FN=FN+1 **end if**
else
if(d_j is a positive data point) **then** TP=TP+1
else FP=FP+1 **end if**
end if
end for
return (TP,TN.FP.FN**)**
end

5. EXPERIMENTAL RESULTS AND DISCUSSION

In this study, three datasets were used and analyzed. The first dataset is 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, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses. In this database, 241 (65.5%) records are malignant and 458 (34.5%) records are benign.

The second dataset is 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 (typical and atypical angina, non-angina pain, and asymptomatic), resting blood pressure, serum cholesterol, resting electrocardiographic results (normal, abnormal, probable), maximum heart rate, indicator of exercise induced angina, ST depression, slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy, and the main criterion that physicians use to determine the diagnosis of heart disease which is the stenosis 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) do not have heart disease.

The third dataset is acquired from database of surgical patients. They have been collected from more than one server computer of Egyptian hospitals. There are 446 records in this database. Each record has 15 features believed to be a good indicator for the infections. Those features include age, gender, clinical department name, type of operation, operation risk index, health degree of patient (from 1 to 5), actual duration time of operation, ideal duration time of operation, wound class (none, mild, moderate, severe), pre- and post-operative time of staying sick, the period between first dose of antibiotic and starting the operation, patient temperature during the operation, infection index (non-infected, infected), and name of organism that caused infection. In this database, 101 records (patients) have infection and 345 records (patients) have no infection.

All the experiments were simulated on AMD Phenom 9550 Quad Core 2.2 GHz workstation with 4 gigabytes of memory running on Windows Server 2003. All calculations and algorithms were implemented on ORACLE 10G. All programs were written in Java language. We divided the data into training and testing phases, in test stage, 5-fold cross validation method was applied. The performance of both classification and clustering in the training phase was used to identify the best parameter setting that can be used, and it was tested in the testing phase. We tabulated the performance for each dataset in Tables 1, 2, and 3, excluding any periods in the classification accuracy that are outside the optimal accuracy region of the solution. We show that the smallest squared error and the smallest information entropy that aim to finding the optimal clustering data [28,25], lead to clustering the dataset into a number of classes that did not hit the highest classification accuracy in training and testing phases. Also, the highest overall accuracy of training phase did not yield the highest classification accuracy in testing phase. On the other hand, the highest Matthew's correlation coefficient in the training phase hit the highest classification accuracy in training and testing phases of each dataset. Therefore, our proposed model depends on MCC as a distinct metric in deciding the optimal number of classes.

Table 1. Training and testing performances using the square error, the information entropy and the percent sensitivity, specificity and overall accuracy and MCC of the proposed approach on the computed clusters of the breast cancer on WDBC database.

Number of Classes	Square Error	Entropy	Training Data				Testing Data			
			Sens.	Spec.	Accu.	MCC	Sens.	Spec.	Accu.	MCC
2	281.36	0.0756	97.85	91.32	95.61	94.61	97.64	88.88	92.31	82.91
3	269.68	0.1435	98.12	92.85	96.31	94.61	97.64	88.88	94.61	88.03
4	261.87	0.1825	97.58	92.85	95.95	95.38	97.64	91.11	93.84	86.32
5	240.99	0.2481	97.05	95.91	96.66	96.92	97.64	95.55	93.84	86.32

6	245.08	0.2058	97.31	94.89	96.48	96.15	97.64	93.33	96.92	93.20
7	247.72	0.2419	97.05	96.42	96.83	96.92	97.64	95.55	96.92	93.20
8	238.22	0.2433	96.78	95.91	96.48	96.15	97.64	93.33	94.61	88.03
9	240.01	0.2453	96.24	97.95	96.83	96.92	97.64	95.55	96.92	93.20
10	235.12	0.2183	96.78	97.95	97.18	96.92	97.64	95.55	96.92	93.20
11	227.93	0.2728	96.78	97.95	97.18	96.92	97.64	95.55	96.92	93.20
12	218.85	0.2707	96.78	97.44	97.01	96.92	97.64	95.55	94.61	88.03
13	212.15	0.2768	97.58	96.42	97.18	96.92	97.64	95.55	96.92	93.20
14	**217.48**	**0.2144**	**97.31**	**98.97**	**97.89**	**95.42**	**97.64**	**97.77**	**97.69**	**94.94**
15	215.77	0.3012	96.24	97.95	96.83	96.92	97.64	97.77	97.69	94.94
16	214.57	0.2975	96.24	97.95	96.83	96.92	97.64	95.55	97.69	94.94
17	211.35	0.2993	95.97	97.95	96.66	96.92	97.64	95.55	97.69	94.94
18	207.98	0.2936	96.78	97.44	97.01	96.92	97.64	95.55	97.69	94.94
19	207.06	0.2975	96.78	97.44	97.01	96.92	97.64	95.55	97.69	94.94
20	189.39	0.2976	96.78	97.95	97.18	96.92	97.64	95.55	97.69	94.94

Table 2. Training and testing performances using the square error, the information entropy and the percent sensitivity, specificity and overall accuracy and MCC of the proposed approach on the computed clusters of the Cleveland heart disease database.

Number of Classes	Square Error	Entropy	Training Data				Testing Data			
			Sens.	Spec.	Accu.	MCC	Sens.	Spec.	Accu.	MCC
2	235.36	0.2265	85.92	89.13	87.49	73.89	82.75	90.90	85.90	72.00
3	**212.55**	**0.2798**	**86.08**	**88.80**	**87.50**	**74.94**	**82.85**	**90.90**	**85.96**	**72.10**
4	200.31	0.2970	73.20	94.34	82.83	71.83	74.28	90.90	80.70	63.48
5	207.14	0.3073	49.22	97.91	75.41	52.11	42.85	100	64.91	47.38
6	186.05	0.3089	62.81	95.82	78.50	52.47	48.57	100	68.42	51.68
7	184.65	0.3073	62.68	96.39	78.82	55.41	45.71	100	66.66	49.52
8	177.32	0.3085	60.23	97.43	78.91	55.36	54.28	95.45	70.17	50.73
9	171.08	0.3095	56.47	99.04	76.66	53.66	45.7	100	66.66	49.52
10	181.32	0.3020	60.21	96.30	79.33	54.50	51.42	100	70.17	53.86

Table 3. Training and testing performances using the square error, the information entropy and the percent sensitivity, specificity and overall accuracy and MCC of the proposed approach on the computed clusters of the surgical patient's database.

Number of Classes	Square Error	Entropy	Training Data				Testing Data			
			Sens.	Spec.	Accu.	MCC	Sens.	Spec.	Accu.	MCC
2	223.60	0.2637	97.13	56.45	91.47	60.95	97.88	60.00	95.97	57.88
3	198.32	0.3062	94.01	72.58	91.03	64.08	97.88	60.00	95.97	57.88
4	191.36	0.3037	98.16	56.45	92.34	64.68	94.74	66.66	89.36	64.29
5	**181.84**	**0.3142**	**95.08**	**71.69**	**91.97**	**65.74**	**94.74**	**66.66**	**89.36**	**64.29**
6	175.56	0.3068	79.24	84.10	83.45	60.12	78.94	88.88	80.85	56.32
7	166.44	0.3197	77.35	84.97	83.95	56.03	76.31	88.88	78.72	53.39
8	165.04	0.3230	75.43	86.79	76.94	54.97	73.68	88.88	76.59	50.64
9	159.79	0.3281	66.47	88.67	69.42	40.08	63.15	88.88	41.04	38.12
10	154.85	0.3322	52.75	94.23	59.06	38.87	44.73	100.0	55.31	36.63

OCFSM depicted in Algorithm 1 requires the setting of the minimal ($c_{\min}$) and maximal ($c_{\max}$) number of classes. In our experiments, we used $c_{\min}$ = 2 and $c_{\max}$ = 20. The optimal number of classes for each dataset is reported in Table 4.

The results show that the WDBC database has several different pathological types of breast tumor disease. The reason for the multiplicity of the pathological types in breast tumors is to point to the presence of many different subsets of features that indicate the disease. But the heart disease has a main feature

for the diagnosis process, which is the stenosis of any major blood vessel. Therefore, Cleveland Heart Disease Database has one pathological type of disease. The surgical patient's database, used in this work, has few types of diseases, and the most influential features in this database leading to infection are the operation risk index, actual duration of the operation, and the ideal duration of the operation. Our proposed model OCSFM is enhanced the classification models behavior, which taken into its consideration during clustering process to representative classes, and compute the optimal number of classes the arriving to the highest MCC in both training and testing phases.

This enhancement can be appeared by sensitive rates comparing with NaïveBayes, Linear SVM, Polykernal SVM, ANN, and SFM models in Tables 5, 6, and 7. This enhancement is reflected on the diagnosis accuracy by helping the pathologist to better detect the type of tumor (benign or malignant), leading to the avoidance of disease complications such as complications of chemotherapy and/or radiotherapy, and the need to undergoing mastectomy. The important improvement is helping physicians to better detect heart diseases and the minimization of post-operative infection. This can be achieved through the avoidance of complications of the disease, and increasing the chances of successful treatment.

Table 4. Optimal number of classes for each dataset using OCSFM.

Dataset	Optimal clusters	Negative clusters (have disease)
Wisconsin breast cancer Diagnostic	14	10
Cleveland Heart Disease	3	1
Surgical patient's	5	2

Table 5. Training and Testing Performance in % sensitivity, specificity, overall accuracy and MCC of NaïveBayes, Linear SVM, Polykernal SVM, ANN, SFM, 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**

Table 6. Training and Testing Performance in % sensitivity, specificity, overall accuracy and MCC of NaïveBayes, Linear SVM, Polykernal SVM, ANN, SFM, OCSFM approaches for Diagnosis in 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**

Table 7 Training and Testing Performance in % sensitivity, specificity, overall accuracy and MCC of NaïveBayes, Linear SVM, Polykernal SVM, ANN, SFM, OCSFM approaches for Diagnosis the infection in 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**

6. CONCLUSIONS

New hybrid model has been proposed and applied to the task of finding an optimal number of different pathological types and its complications for many diseases. These include heart diseases, breast cancer, and post-operative infections. This model extracts an optimal partitioning representative and the most significant features for each pathological type. This approach employs a combination of clustering, selected features and classification concepts. Results have indicated that our proposed approach can minimize noise data points, smallest intra-class distance, and greatest inter-class distance for all classes. It also finds the optimal selection of features in order to have the maximum number of samples correctly classified, thus yielding the highest classification accuracy possible. Here, after applying our intelligent model, the powerful accuracy evaluation criterion MCC of machine learning methods have been improved compared with NaïveBayes, Linear SVM, Polykernal SVM, ANN, and SFM approaches and is reflected on the accuracy of the diagnosis process.

7. REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[16] Blatt, M., Wiseman, S., and Domany, E., 1996. Super-paramagnetic clustering of data. Phys Rev Lett. 76, (3251-3254):29-76.

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

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

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

[20] Jiang, D., Pei, J., and Zhang, A., 2003. 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.

[21] Hinneburg, A., and Keim, D., 1998. 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.

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

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

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

[25] Romdhane, L., Shili, H., and Ayeb, B., 2009. Mining microarray gene expression data with unsupervised possibilistic clustering and proximity graphs. Appl Intell. 10.1007/s (10489-009):0161-3.

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

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

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

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