SVM Based Clinical Decision Support System For Accurate Diagnosis Of Chronic Obstructive Pulmonary Disease

[*]Sandhya Joshi and [1]Hanumanthachar Joshi
[*]*Department of Studies in Computer Science,*
Pooja Bhagavat Memorial Mahajana PG Centre, Mysore.

[1]*Department of PG Studies and Research,*
Sarada Vilas College of Pharmacy, Mysore.

Abstract

Chronic Obstructive Pulmonary Disease (COPD) is a major public health problem worldwide. Its prevalence, morbidity, and mortality rates are increasing and it is now the fourth leading cause of death worldwide. The study comprised of 495 patients suffering from COPD were identified through hospital and primary care registers based on the GOLD criteria for the classification of various stages of COPD. An attempt was made to develop a Clinical Decision Support System (CDSS) for the diagnosis of the various stages of the COPD by considering around 22 features. The study consisted of selection of feature subsets, fishers score method was employed for identifying the most prominent features. Features were fed to the Support Vector Machines classifier and Particle Swarm Optimization method was used for optimizing the parameters to design a clinical decision support system (CDSS) for the diagnosis of the various stages of COPD. The results were compared with SVM based on the grid search technique and principle component analysis (PCA-Grid-SVM) in terms of their classification accuracy. In addition, classification accuracy of the present study was compared to the previous studies and the proposed system achieved the highest classification accuracy with 96.75% when compared to the existing methods. Hence, the proposed CDSS can be used as a supporting tool and can be enormous help in assisting the physicians to make the accurate diagnosis on the patients with COPD along with Spirometry.

Keywords: SVM classifier, Feature selection Particle Swarm Optimization, Principle component Analysis

1. Introduction

COPD is an inflammatory lung disease characterized by a permanent blockage of airflow from the lungs. The primary cause of COPD is tobacco smoke (through smoking or second-hand smoke). The disease is widely under-diagnosed, although it is a life-threatening lung disease, which is not fully reversible. COPD is one of the leading non-infectious diseases in the world. The World Health Organization (2009) estimates that about 210 million people worldwide have COPD [1]. It is currently the fourth leading cause of death, and they predict that by 2030 it will have become the third leading cause of death. The most important cause of COPD is smoking. Not only active smoking, but also passive exposure to smoking, air pollution and occupational chemicals contribute to a higher risk for COPD in both high-income and low-income countries. Because of the slow progression of the disease, COPD is frequently not diagnosed until after the age of 40. It's for this reason that the overall prevalence is low before the age of 40 and increases with age. Due to increase in smoking habits among women and a greater risk of exposure to air pollution, prevalence amongst men and women is almost equal today (World Health Organization [WHO], 2009). COPD is chronic and progressive and there are no remedies to cure COPD. Existing treatments can only slow down the progression of the disease, preventing symptoms or by reducing the severity of existing symptoms and associated complications.

The diagnosis of COPD is based on three different parts, symptoms, assessment of lung function and the evaluation of the responses to inhaled pharmacological agents. Although these tests are generally considered informative, they are time-consuming and strongly dependent on the personnel's professional experience. Under diagnosis of COPD is a big problem [2, 3, 4, 5, 6] and may be caused by patient delay or sometimes doctors delay. One reason for under-diagnosis is due to the delay between symptoms develop and first consultation. Many of the symptoms of COPD are associated as being normal for a smoker, such as morning cough and sputum production. People may blame themselves for these symptoms and do not seek a doctor until they feel ill. Smokers rarely seek medical advice specifically for cough [7].

Spirometry is a low cost and relatively non-invasive approach but it requires subject effort, and repeated exhalation manoeuvres can be difficult for breathless people. Spirometry also requires trained staff and is affected by factors such as technique, age, height, gender and ethnic origin; its accuracy also decreases in young or very elderly subjects [8]. A lack of local reference values for most of the world's population, disparity between symptoms and lung function and negative screening studies [9] highlight the limitations of spirometry.

In India, normally COPD diagnosis is based on spirometry. The use of GOLD criteria (for spirometry) will underestimate the number of patients with COPD in persons 50 years and younger [8]. Data from National Health and Nutrition Examination Survey (NHANES3) and Health survey of England (HSE) confirm that using FEV1/FVC<70% to define obstruction will cause 14% under-classification in those 50 years and younger [10]. In order to simplify the diagnosis of COPD GOLD recommends a FEV1/FVC threshold of 70% regardless of age. The FEV1/FVC ratio falls with age [11, 12]. The use of a fixed cut-off point for defining COPD becomes more inaccurate with increasing age [8]. The criteria for diagnosing chronic obstructive pulmonary disease (COPD) are still under debate [13, 14]. Van Berkel et al proposed a set of six Volatile Organic Compounds (VOC's) that could differentiate COPD patients from controls with high sensitivity and specificity [15] and Fens et al suggested exhaled breath profiling discriminates COPD from asthma albeit with some overlap in profiles between COPD and asymptomatic smokers when using an electronic nose [16]. However, these studies are small, have not validated smoking status, and have not considered the age, body mass index (BMI), arterial Oxygen Partial Pressure (PaO2) arterial Carbon Dioxide Partial Pressure(PaCO2), and gender of subjects—all of which may affect VOCs [17-19]. Kishkel et al showed how important these confounders could be when apparent differences of exhalation profiles between lung cancer and non-cancer patients did not persist after these confounding variables were taken into account [20]. The same can be considered true for COPD. Although the risk of under-diagnosis of COPD is important, the issue of over diagnosis and over-treatment in the elderly also needs discussion. Therefore in the current study an attempt was made to develop a Clinical Decision Support System for the diagnosis of COPD. An appropriate CDSS can highly increase diagnosis accuracy, improve healthcare quality, and reduce cost.

In the present study, feature subsets are obtained by using Fisher score method CDSS integrating support vector machines and particle swarm optimization to construct the model for the diagnosis of patients with mild, moderate, severe and very severe stages of COPD. This diagnosis system aims to maximize the generalization capability of SVM for the classification of various stages of COPD.

Stages	GOLD classification (%)	Mean ± SD
Mild Stage I	FEV1/FVC<0.7; FEV1>80% (predicted)	9.8
Moderate Stage II	FEV1/FVC<0.7; 50%< FEV1< 80% (predicted)	19.5
Severe Stage III	FEV1/FVC<0.7; 30%< FEV1< 50% (predicted)	28.4
Very Severe Stage IV	FEV1/FVC<0.7 FEV1< 30% (predicted)	52.2

Table 1. GOLD Classification of COPD by Severity

2. Datasets and problem definition

In the current study COPD patients were identified through hospital and primary care registers from different parts of northern Karnataka. All were deemed stable by a respiratory clinician and none reported worsening symptoms within six weeks of testing. All were prescribed optimal medication [21]. Participant's undergone questionnaires for socio-demographic data like age [22], Sex [23,24], BMI [25,26], smoking status and any illnesses including current/recent symptoms then performed dry wedge Spirometry for breath analysis of various VOC's were recorded. Smoking status was validated using exhaled carbon monoxide (CO). Since COPD has impacts beyond the lung [23], the prognostic impact of both pulmonary and extra-pulmonary factors has been investigated. Parameters of lung function and particularly forced expiratory volume in 1 second (FEV1) [24], arterial blood gases [27, 28], arterial Oxygen Partial Pressure, PaCO2: arterial Carbon Dioxide Partial Pressures. pulmonary function tests were used to assess the severity of disease in COPD including Spirometry.

The study comprised of 495 patients with COPD. COPD was observed in differing severities which are described according to internationally agreed guidelines from the Global initiative for chronic obstructive Lung Disease (GOLD criteria). Based on the severity of the lung impairment the patients were classified into four stages as mentioned in Table 1. In this study, for every patient FEV1, FVC, FEV1/FVC ratio, Total Lung Capacity (TLC), Functional Residual Capacity (FRC), Residual Volume (RV) and Inspiratory Capacity to TLC ratio (IC/TLC) were recorded. The main objective of the study was to design a CDSS integrating support vector machines and particle swarm optimization to construct the model for the diagnosis of patients with mild moderate, severe, very severe stage of COPD. The detailed characteristics of the study population are shown in Table 2.

Table 2. Characteristics of the Study population

Features	*Attributes*	*Mean ± SD*
f_1	Age (years)	61.9±9.7
f_2	Sex (%) Male	62.9
	Female	37.1
f_3	BMI (kg/m2)	24.2±5.3
f_4	Smoking (%)	15.6
	Pack-years	26.8
	Exacerbations (%)	42.8
	0-1/year	33.5
	2-4/year	15.6
	>4/year	
f_5	FEV1 (%predicted)	37±18.1
f_6	FVC (%predicted)	34.7±12.3
f_7	FEV1/FVC	84.5±22.4
f_8	TLC (%predicted)	124.9±18.4
f_9	IC/TLC (%)	26.9±8.8
f_{10}	RV (%predicted)	205.1±58.4
f_{11}	FRC (% predicted)	172.7±38.5
f_{12}	DLco (%predicted)	40.8±18
f_{13}	PaO2 (KPa)	9.4±1.4
f_{14}	PaCO2 (KPa)	5.2±0.9
	VOC's frequency in (%)	
f_{15}	Benzaldehyde	100
f_{16}	Toluene	89.3
f_{17}	Benzene	98.6
f_{18}	Sulphur dioxide Benz	76.8
f_{19}	Hexanal	91.5
f_{20}	Isoprene	93.2
f_{21}	1-Heptene	93.2
f_{22}	Acetic acid	86.4

(BMI: Body Mass Index; FEV1: Forced Expiratory Volume in 1second; FVC: Forced Vital Capacity; GOLD: Global Initiative for Obstructive Lung disease; TLC: Total Lung Capacity; RV: Residual Volume; IC: Inspiratory Capacity; FRC, Functional Residual Capacity; DLco; Carbon Monoxide Transfer Factor; PaO2: arterial Oxygen Partial Pressure; PaCO2: arterial Carbon Dioxide Partial Pressure)

3. Data Analysis

3.1 Feature Selection

Feature selection is a process by which a sample in the measurement space is described by a finite and usually smaller set of number classed features. The objective of Feature Selection is to determine a minimal feature subset from a problem domain whilst retaining a suitably high accuracy in representing the original features. The usefulness of a feature subset can be determined by both its relevance and redundancy [29].

3.1.1. Fisher Score

The key idea of Fisher score is to find a subset of features, such that in the data space spanned by the selected features, the distances between data points in different classes are as large as possible, while the distances between data points in the same class are as small as possible. Fisher Score [30] is a supervised feature selection method for determining the most relevant features for classification. It selects a good feature by the score that is measured by its discriminative power defined by Fisher's Criterion. Given data with labels $\{x_i, y_i\}$, y_i Є $\{1,2..c\}$. Let n_i denote the number of samples in class i and μ_i denote the mean value of class i. Let μ denote global mean value, σ_i^2 denote the variance of class i. The Fisher Score (F) criterion is computed as follows:

$$F= \frac{\sum_{i=1}^{c} n_i(\mu_i-\mu)2}{\sum_{i=1}^{c} n_i(\sigma_i)2}$$

Fisher Score directly measures the value F for each feature. A feature will have a high score if it has high between class scatter and low within class scatter. In the current study we used fisher score method for determining the most predominant attributes for the classification of various stages of COPD.

3.2 Support Vector Machine and Particle Swarm Optimization

The support vector machine (SVM) is a supervised learning method widely used for classification. It is a powerful methodology for solving problems in nonlinear classification, function estimation, and density estimation, leading to many applications including image interpretation, data mining, biometric authentication, biotechnological investigation, and clinical diagnosis [31-34]. This section gives a brief description on SVM. For more details, one can refer to [35], which give complete descriptions of the SVM theory. Let us consider a binary classification task$\{x_i, y_i\}$; i =1, . . . *l*; $y_i \in \{-1, 1\}$, $x_i \in R^d$, where x_i are data points and y_i are corresponding labels. They are separated with a hyper plane given by $w^T x + b = 0$, where *w* is a d-dimensional coefficient vector which is normal to the hyperplane and *b* is the offset from the origin. The linear SVM finds an optimal separating margin by solving the following optimization task:

$$\text{Minimize } g(w,\xi)=\tfrac{1}{2}\|w\|^2 + m\textstyle\sum_{i=1}^{n} \xi_i \qquad (1)$$

$$\text{Subject to : } y_i(w^T x_i + b) \geq 1\text{-}\xi_i, \xi_i \geq 0 \quad (2)$$

Where *m* is a penalty value, ξi is the positive slack variables. This primal problem can be reduced to the Lagarangian dual problem by introducing Lagrangian multipliers *αi*. According to the Karush Kuhn–Tucker (KKT) condition, we can get the optimal solution αi. If αi>0, the corresponding data points are called SVs. Afterwards, we can get the optimal hyperplane parameters w and b. Then the linear discriminant function can be given by

$$g(x)=\text{sgn}(\textstyle\sum_{i=1}^{n} \alpha_i y_i\, k(x_i, x) + b) \qquad (3)$$

In order to make the linear learning machine work well in non-linear cases, the original input space can be mapped into some higher-dimensional feature space via a mapping function. With this mapping, x^T_i x in the input space can be represented as the form of $\emptyset(x_i)^T \emptyset(x)$ in the feature space. The functional form of the mapping $\phi(x_i)$ does not need to be known since it is implicitly defined by one selected kernel: $k(x_i,x_j) = \phi(x_i)^T/(x_j)$. Two most widely used kernels in SVM are the polynomial kernel and the Gaussian kernel (or Radial-Basis function, RBF.

Particle swarm optimization algorithm seeks to explore the search space by a population of individuals or particles. Each particle represents a single solution with a velocity which is dynamically adjusted according to its own experience and that of its neighboring companions. The population of particles is updated based on each particle's previous best performance and the best particle in the population. Particle swarm optimization combines local search with global search for balancing the exploration and exploitation [36].

Considering a d-dimensional search space, the ith particle is represented as $\overrightarrow{X_i}$ =$(x_{i,1} x_{i,2}, \dots x_{i,d})$, and its according velocity is represented as$_{\overrightarrow{V_i}}$ =$(v_{i,1} v_{i,2}, \dots v_{i,d})$,. The best previous position of the ith particle that gives the best fitness value is represented as $\overrightarrow{P_i}$ =$(p_{i,1} p_{i,2}, \dots p_{i,d})$,. The best particle among all the particles in the population is represented as $\overrightarrow{P_g}$ =$(p_{g,1} p_{g,2}, \dots p_{g,d})$, In every iteration, each particle updates its position and velocity according to the two best values.

In order to reduce the dependence of the search process on the hard bounds of the velocity, the concept of an inertia weight w was introduced in the Particle swarm optimization algorithm [37]. The velocity and position are updated as follows:

$$V_{i,j}^{n+1}=w\ \text{X} v_{i,j}^{n}+a_1 x r_1 (p_{i,j}^{n}-x_{i,j}^{n})+ a_2 x r_2 (p_{g,j}^{n}-x_{i,j}^{n}) \quad (4)$$

$$x_{i,j}^{n+1}=x_{i,j}^{n}+V_{i,j}^{n+1},\ j = 1,2, \dots .d \qquad (5)$$

where a_1 and a_2 are acceleration coefficients, which define the magnitude of the influences on the particles velocity in the directions of the personal and the global optima, respectively. To better balance the search space between the global exploration and local exploitation, Time-Varying Acceleration Coefficients (TVAC) has been introduced in [38]. This concept will be adopted in this study to ensure the better search capability for the solutions. The core idea of TVAC is that a_1 decreases from its initial value of a_{1i} to a_{1f}, while a_2 increases from a_{2i} to a_{2f} using the following equations as in. TVAC can be mathematically represented as follows:

$$a_1= (a_{1f}-a_{1i})\frac{t}{t_{max}}+a_{1i} \qquad (6)$$

$$a_2= (a_{2f}-a_{2i})\frac{t}{t_{max}}+a_{2i} \qquad (7)$$

where a_{1f}, a_{1i}, a_{2f} and a_{2i} are constants, *t* is the current iteration of the algorithm and t_{max} is the maximum number of iterations. In addition, r_1 and r_2 in Eq. 8 are random numbers, generated uniformly in the range [0, 1]. The velocity $v_{i,j}$ is restricted to the range $[-v_{max}, v_{max}]$, in order to prevent the particles from flying out of the solution space. The inertia weight *w*, which is used to balance the global exploration and local exploitation, a large inertia weight facilitates the global search, while a small inertia weight facilitates the local search. In order to reduce the weight over the iterations allowing the algorithm to exploit some specific areas, the inertia weight *w* is updated according to the following equation:

$$W = w_{min}+(w_{max}-w_{min})\frac{(t_{max}-t)}{t_{max}} \qquad (8)$$

Where w_{max}, w_{min} are the predefined maximum and minimum values of the inertia weight w, t is the current iteration of the algorithm and t_{max} is the maximum number of iterations. Usually the value of w is varied between 0.9 and 0.4. Eq. 9 is also known as the Time-Varying Inertia Weight (TVIW). It has been shown to significantly improve the performance of PSO [39], since TVIW makes PSO have more global search ability at the beginning of the run and have more local search ability near the end of the run.

4. CDSS designed with PSO and SVM for the diagnosis of COPD

As discussed in previous section the aim of this clinical decision support system is to maximize the generalization capability of SVM for chronic obstructive pulmonary disease diagnosis. In order to achieve this goal, we started with selection of feature subsets using Fishers score method. In the second stage, different feature subsets are fed into the SVM classifier for training an optimal model, in the meanwhile the model parameters of SVM are optimized using particle swarm optimization algorithm. At the end, SVM model conducts the diagnostic tasks using the most discriminative feature subset and the optimal parameters. The main component of the proposed system is the parameters optimization process.

The overall optimization procedure begins with construction of parameter optimization, by encoding the particle with the dimensions. The first two dimensions *a* and γ, which are the model parameters of SVM. The third one is the linear weight of the sub-objective functions, and then the individuals of the population with random numbers are initialized and in the meanwhile, specify the PSO parameters including the lower and upper bounds of the velocity, the size of particles, the number of iterations, etc. Then train SVM with the initialized parameters i.e. the initial particle and feasible random numbers. Then the particle with high classification accuracy can produce a high fitness value. Moreover, the particle with smaller number of Support vectors can achieve higher classification accuracy, since the number of Support Vectors is proportional to the generalization error of the SVM classifier. So, both of them are taken into account to design the objective function.

The fitness value is calculated according to the following function:

$$\begin{cases} f1 = avgacc = \dfrac{(\sum_{i}^{K} Test\_Accuracyy_{i})}{K} \\ f2 = (1 - \dfrac{nsv}{m}) \\ f = \alpha X f_1 + (1-\alpha) X f2 \end{cases} \qquad (9)$$

Where variable *avgacc* in the first sub-objective function f_1 represents the average test accuracy achieved by the SVM classifier via K-fold CV, where K=5. Note that here the 5-fold CV is employed to do the model selection that is different from the outer loop of 10-fold CV, which is used to do the performance estimation. *nsv* and *m* in the second sub-objective function f_2 indicates the number of support vectors and training data, respectively. The weighted summation of the two sub-objective functions is selected as the final objective function. In *f*, variable α is the weight for SVM classification accuracy and 1-α indicates the weight for the number of supportive vectors. The weight can be adjusted to a proper value depends on the importance of the sub-objective function. Eq.9 represents the classification accuracy and the numbers of supportive vectors have different significance to the classification performance. Generally, the weight is set to be constant value according to the specified problem at hand. In other words, setting the weight is problem dependent. Here we take into account α in optimization for evolving the optimal values. In this way, the weight of each sub-objective can be adaptively specified. After the fitness value was obtained, the global optimal fitness was saved as *gfit*, personal optimal fitness as *pfit*, global optimal particle as *gbest* and personal optimal particle as *pbest*. Radial-basis function, $k(x_i,x_j)$ increases the number of iteration. Whereas kernel function *g(x)* increases the number of population. Update the position and velocity of *a* and γ in each particle according to Eqs. 4 and 5. Once it is updated then train the SVM model and calculate the fitness value of each particle according to Eq. 9. Update the personal optimal fitness (*pfit*) and personal optimal position (*pbest*) by comparing the current fitness value with the *pfit* stored in the memory. If the current fitness is dominated by the *pfit* stored in the memory, then keep the *pfit* and *pbest* in the memory; otherwise, replace the *pfit* and *pbest* in the memory with the current fitness value and particle position. If the size of the population is reached, then update the global optimal fitness (*gfit*) and global optimal particle (*gbest*) by comparing the *gfit* with the optimal *pfit* from the whole population, If the current optimal *pfit* is dominated by the gfit stored in the memory, then keep the *gfit* and *gbest* in the memory; otherwise, replace the *gfit* and *gbest* in the memory with the current optimal *pfit* and the optimal *pbest* from the whole population otherwise increase the number of population and update position and velocity, procedure repeates till it the iteration number reaches the maximum number of iterations then finally we get the optimal (*a*, γ) and the feature subset from the best particle (*gbest*).

Table 3. The Relative Important Features Obtained from Fisher Score

Feature Numbers	Features (Attributes)	Fisher Score
f_7	FEV1/FVC	3.25
f_5	FEV1	2.75
f_6	FVC	2.23
f_9	IC/TLC	2.05
f_{15}	VOC's (Benzaldehyde)	1.70
f_{12}	DLco	1.58
f_{13}	PaO2 (KPa)	1.32

5. Experimental procedure

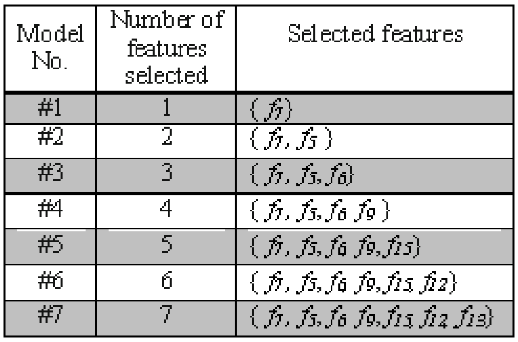
The sample consisted of initial visits of 495 subjects seen as patients suffering from COPD in different stages. The Physicians evaluated the patients based on the GOLD criteria for the classification of various stages of COPD along with socio-demographic data and spirometry for breath analysis and description of around nine volatile organic compounds was recorded. Around 22 features were observed and all of them were continuous, the details of the whole dataset are mentioned in the Table 2. The procedure begins with scaling all the attributes so that they lie in a suitable range. Usually, the data could be normalized by scaling them into the interval of [−1, 1] according to the Eq. 10, where x is the original value, x' is the scaled value, max_a is the maximum value of feature a, and min_a is the minimum value of feature a.

$$x' = \left(\frac{x - min_a}{max_a - min_a}\right) * 2 - 1 \qquad (10)$$

In order to validate the results, 10 fold Cross Validation is used to evaluate the classification accuracy. This study set K as 10, i.e., the data was divided into ten subsets. Each time, one of the ten subsets is used as the test set and the other nine subsets are put together to form a training set. Then the average error across all ten trials is computed. The advantage of this method is that all of the test sets are independent and the reliability of the results could be improved. In order to ensure the same class distribution in the subset, the data is split via stratified sampling in which the sample proportion in each data subset is the same as that in the population. Empirical studies showed that stratified CV tends to generate comparison results with lower bias and lower variance when compared to regular k-fold CV [40]. Note that only one repetition of the 10-fold CV will not generate enough classification accuracies for comparison. Because of the arbitrariness of partition of the data set, the predicted accuracy of a model at each iteration is not necessarily the same. To evaluate accurately the performance of the data sets, the 10-fold CV shall be repeated 5 times and then averaged the results.

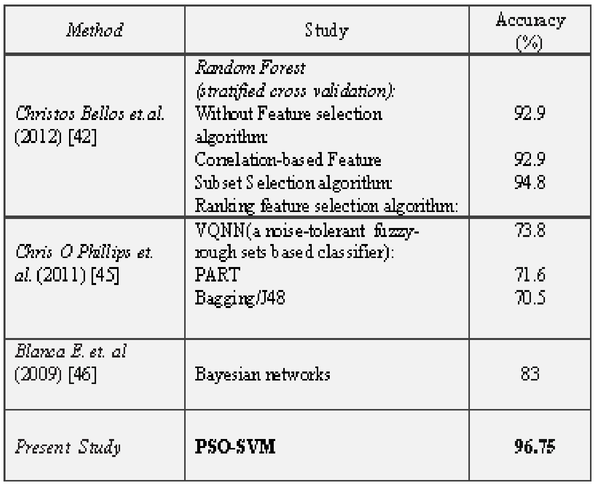
The CDSS model for COPD was implemented using MATLAB platform. The Matlab short for matrix laboratory is a useful tool which provides a many important data mining tools which include the Neural Network, Support Vector Machines etc. For SVM, LIBSVM implementation was utilized, which is originally developed by Chang and Lin [41]. PSO algorithm was used from [42], The number of iterations and particles was set to 100 and 25, respectively. As indicated in [43] the searching ranges for a and γ were as follows: a ε [10^ (–2), 10^ (2)] and γ ε [10^ (–2), 10^ (2)]. v_{max} is set about 60% of the dynamic range of the variable on each dimension. As suggested in [43], a_{1i}, a_{1f}, a_{2i} and a_{2f} were set as follows: a_{1i}=2.5, a_{1f} = 0.5, a_{2i} = 0.5, a_{2f} = 2.5. Where as w_{max} *and* w_{min} were set to 0.9 and 0.4 respectively. The same data set was used to build Grid-SVM model. For Grid-SVM, the range of the related parameters a and γ were varied between a =$\{2^{-5}, 2^{-3}, \ldots 2^{-15}\}$ and γ =$\{2^{-15}, 2^{-13}, \ldots 2^{1}\}$ the grid search technique [44] was employed using 5-fold CV to find out the optimal parameter values of RBF kernel function.The freeware LIBSVM, a library for SVM, was adopted for integration.

Table 4. The Seven Feature Subsets based on FS.



Model No.	Number of features selected	Selected features
#1	1	{ f_7}
#2	2	{ f_7, f_5 }
#3	3	{ f_7, f_5, f_6}
#4	4	{ f_7, f_5, f_6 f_9 }
#5	5	{ f_7, f_5, f_6 f_9, f_{15}}
#6	6	{ f_7, f_5, f_6 f_9, f_{15} f_{12}}
#7	7	{ f_7, f_5, f_6 f_9, f_{15} f_{12} f_{13}}

Table 6. Classification Accuracies Obtained with other Methods



Method	Study	Accuracy (%)
Christos Bellos et.al. (2012) [42]	*Random Forest (stratified cross validation):* Without Feature selection algorithm: Correlation-based Feature Subset Selection algorithm: Ranking feature selection algorithm:	 92.9 92.9 94.8
Chris O Phillips et. al. (2011) [45]	VQNN(a noise-tolerant fuzzy-rough sets based classifier): PART Bagging/J48	73.8 71.6 70.5
Blanca E. et. al (2009) [46]	Bayesian networks	83
Present Study	**PSO-SVM**	**96.75**

6. Results and discussion:

In the present study Fisher score method was employed and seven prominent features were selected (Table 3). It can be observed that the degree of importance of each feature from high to low i.e., f_7 , f_5, f_6, f_9, $f_{15,}$ f_{12} and $_,f_{13}$. The ratio of forced expiratory volume to the forced vital capacity in 1 sec is ranked with a score of 3.25. Seven models were constructed with different number of features to obtain the SVM classification models. Seven models with different features subsets based on fishers score (Table 4). The classification results obtained over seven runs of ten fold cross validation are represented in Table 5.

Table 5. Classification accuracies obtained from FS-PSO-SVM Method

Model	Classification Accuracy (%)			
	Mean	*SD*	*Max*	*Min*
#1	91.45	0.32	92.19	90.02
#2	91.95	0.45	92.45	91.21
#3	93.84	0.59	94.39	92.75
#4	95.02	0.83	96.53	94.56
#5	*96.75*	0.44	*96.97*	95.23
#6	96.57	0.52	96.81	95.16
#7	96.32	0.76	97.12	95.26

It can be observed from the Table 5 that the best results 96.75% mean classification and 97.97% maximum classification were obtained with the model #5. Hence model #5 can be regarded as the best feature subset among seven feature subset based on classification accuracy. By comparing the results of classification accuracy obtained with PSO-SVM, and Fisher Score-PSO-SVM (model #5), it was observed that there has been an improvement in mean classification accuracy by 0.45% and maximum classification accuracy by 0.92% upon using fisher score. Table 6 represents the comparison of various classification accuracies obtained by the previous methods and the current method. It was observed that the proposed CDSS developed by integrating Fishers score-PSO-SVM achieved better performance accuracy when compared with all other available methods proposed in the previous studies.

For understanding the effectiveness of the proposed system an attempt has been made to compare the results of the proposed method with that of Grid-SVM and Principle Component Analysis with Grid-SVM. In PCA all the seven principle components were fed in to the SVM classifier. Table 6 shows various classification results obtained with different number of principle components. The mean classification accuracies were found to be 91.63% to 93.94% and the maximum classification accuracy were 92.43% - 94.76%. A profound increase in mean (0.60%) and maximum (0.65%) of classification accuracy was found in PCA- Grid-SVM model as compared with that of Grid-SVM model (Table 7).

Table 7. Classification Accuracy obtained from PCA-Grid-SVM method

No. of principal components	Classification Accuracy (%)			
	Mean	*SD*	*Max*	*Min*
1	91.63	0.24	92.85	91.02
2	92.08	0.46	92.97	91.61
3	92.56	0.35	93.45	92.11
4	91.68	0.71	92.43	90.54
5	***93.94***	0.68	***94.76***	92.13
6	93.07	0.53	94.42	92.65
7	93.45	0.76	94.16	92.16

The summary of various classification accuracies obtained from PSO-SVM and Grid-SVM is shown in Table 8. It was observed that PSO-SVM has via 5 runs of 10 fold CV proven better performance when compared with Grid-SVM in terms of classification accuracy at the statistical significance level of 5%.

Table 8. The Performance Comparison of PSO- SVM with Grid-SVM (5%)

Five runs of 10-fold CV	PSO-SVM	Grid-SVM
#1 run	96.23	93.43
#2 run	95.89	93.10
#3 run	96.75	93.68
#4 run	96.54	92.86
#5 run	96.13	93.16
Mean ± SD	96.30±0.25	93.24±0.48
For Confidence level α = 5% ; *Paired t*-test *p*-value = 0.028		

This proves that particle swarm optimization approach has better chance of finding the global optimal solution as compared to grid method. The detailed classification results of Fisher score-PSO -SVM and PCA- Grid SVM (using 5 PC's) via 5 runs of 10- fold CV is shown in Table 9. It was observed that results from FS-PSO-SVM were higher than PCA-Grid-SVM by 2.98% and the superiority of FS-PSO-SVM is statistically significant at the level of 10%. The results clearly imply that Fishers score method is more efficient in constructing discriminative feature space for classification than PCA method.

Table 9. The performance comparison of PSO- SVM with Grid-SVM (10%)

Five runs of 10-fold CV	FS-PSO-SVM	PCA-Grid-SVM
#1 run	97.11	94.86
#2 run	96.23	94.64
#3 run	96.28	94.58
#4 run	96.16	93.71
#5 run	97.26	94.62
Mean ± SD	96.98±0.46	94.83±0.82
For Confidence level α = 10% *Paired t*-test *p*-value = 0.0764		

From the above results, it can be observed that the clinical decision support system obtained by integrating FS-PSO-SVM is more appropriate for the diagnosis of COPD as compared to other methods. Hence, the proposed CDSS can be used as a supporting tool and can be enormous help in assisting the physicians to make the accurate diagnosis on the patients with COPD along with Spirometry.

7. References

[1] World Health Statistics, World Health Organization, 2009.

[2] Global initiative for Chronic Obstructive Lung Disease, Global strategy for the diagnosis, management and prevention of chronic obstructive pulmonary disease NHLBI/WHO workshop report. National Institutes of Health, National Heart, Lung, and Blood Institute. 2001; 2005. Report No.: 2701.

[3] A. Johannessen, E. Omenaas, P. Bakke,A. and Gulsvik, "Incidence of GOLD-defined chronic obstructive pulmonary disease in a general adult population", *Int J Tuberc Lung Dis*, Aug 2005, pp. 926-32.

[4] S. Weiss, D. DeMeo and D.S. Postma, "COPD:problems in diagnosis and measurement", *Eur Respir J*, 2003, pp. 4-12.

[5] D. J. Brazzale, A. L. Upward, and J. J. Pretto, "Effects of changing reference values and definition of the normal range on interpretation of spirometry", *Respirology*, 2010, pp. 1098-103.

[6] S. C. Hvidsten, L. Storesund, T. Wentzel-Larsen, A. Gulsvik, and S. Lehmann, "Prevalence and predictors of undiagnosed chronic obstructive pulmonary disease in a Norwegian adult general population", *Clin Respir J* , Jan. 2010, pp. 13-21.

[7] A. H. Morice, G. A. Fontana, A. R. Sovijarvi, M. Pistolesi, K.F. Chung and J. Widdicombe, " The diagnosis and management of chronic cough", *Eur Respir J*, Sep. 2004, pp. 481-92.

[8] J. A. Hardie, A. S. Buist, W. M. Vollmer, I. Ellingsen, P. S. Bakke and O. Morkve, "Risk of over-diagnosis of COPD in asymptomatic elderly never-smokers", *Eur Respir J*, Nov. 2002, pp. 1117-22**.**

[9] NICE guidelines chronic obstructive pulmonary disease. National Institute for Health and Clinical Excellence, 2012.

[10] S.C. Hvidsten, L. Storesund, T. Wentzel-Larsen, A. Gulsvik and S. Lehmann, " Prevalence and predictors of undiagnosed chronic obstructive pulmonary diseasein a Norwegian adult general population", *Clin Respir J*, Jan. 2010 pp. 13-21.

[11] J. L. Hankinson, J. R. Odencrantz and K. B. Fedan, "Spirometric reference values from a sample of the

general U.S. population", *Am J Respir Crit Care Med*, Jan.1999, pp. 179-87**.**

[12] A. Tamimi, D. Serdarevic and N. A. Hanania, "The effects of cigarette smoke on airway inflammation in asthma and COPD: Therapeutic implications", *Respir Med*, Mar. 2012, pp. 319-28.

[13] B. R. Celli, R. J. Halbert, S. Isonaka and B. Schau, "Population impact of different definitions of airway obstruction", *Eur Respir J*, Aug. 2003, pp. 268-73.

[14] ATS/ERS guidelines for COPD. American Thoracic Society and European Respiratory Society, 2012.

[15] Van Berkel J J *et al* 2010 A profile of volatile organic compounds in breath discriminates COPD patients from controls *Respir. Med.* 104 557-63.

[16] N Fens, "Exhaled breath profiling enables discrimination of chronic obstructive pulmonary disease and asthma", *Am. J. Respir. Crit. Care Med.*, 2009, pp. 1076–82.

[17] B. Buszewski, "Human exhaled air analytics: biomarkers of diseases", *Biomed. Chromatogr*, 2007, pp. 553-66.

[18] A. Amann, "Applications of breath gas analysis in medicine", *Int. J. Mass Spectro*, 2009, pp. 227-233.

[19] M. Phillips, "Effect of age on the breath methylated alkane contour, a display of apparent new markers of oxidative stress", *J. Lab. Clin. Med.* 2000, pp. 243-249.

[20] S. Kischkel, "Breath biomarkers for lung cancer detection and assessment of smoking related effects-confounding variables, influence of normalization and statistical algorithms", *Clin. Chim. Acta*, 2010, 1637-44.

[21] Global Strategy for the Diagnosis, Management and Prevention of COPD. Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2010.

[22] N. R. Anthonisen, E. C. Wright and J. E. Hodgkin, "Prognosis in chronic obstructive pulmonary disease", *Am Rev Respir Dis."*, 1986, pp. 14-20.

[23] R. Antonelli Incalzi, L. Fuso, M. De Rosa, F. Forastiere, E. Rapiti, B. Nardecchia and R. Pistelli, "Co-morbidity contributes to predict mortality of patients with chronic obstructive pulmonary disease", *Eur Respir J*, 1997, pp. 2794-2800.

[24] G. Gudmundsson, T. Gislason, E. Lindberg, R. Hallin, C. S. Ulrik, E. Brondum, M. M. Nieminen, T. Aine, P. Bakke and C. Janson, " Mortality in COPD patients discharged from hospital: the role of treatment and co-morbidity", *Respir Res*, 2006, pp. 77- 109.

[25] A. M. Schols, J. Slangen, L. Volovics and E. F. Wouters, " Weight loss is a reversible factor in the prognosis of chronic obstructive pulmonary disease", *Am J Respir Crit Care Med*, 1998, pp. 1791-1797.

[26] E. Chailleux, J. P. Laaban and D. Veale, "Prognostic Value of Nutritional Depletion in Patients with COPD Treated by Long-term Oxygen Therapy: Data From the ANTADIR Observatory", *Chest*, 2003, pp. 1460-1466.

[27] K. H. Groenewegen, A. M. Schols, E. F. Wouters, "Mortality and mortality related factors after hospitalization for acute exacerbation of COPD", *Chest*, 2003, pp. 459-467.

[28] O. A. Yildiz, Z. P. Onen, E. Sen, B. E. Gulbay, K. Kose, S. Saryal and G. Karabiyikoglu, " Predictors of long-term survival in patients with chronic obstructive pulmonary disease", *Saudi medical journal*, 2006: 27(12): 1866-1872.

[29] I. Guyon, and A. Elisseeff, "An introduction to variable and feature selection" *Journal of Machine Learning Research*, 2003, pp. 1157-1182.

[30] Chao-Ton Su and Chien-Hsin Yang, "Feature selection for the SVM: An application to hypertension diagnosis", *Expert Systems with Applications*, 2008, pp. 754–763.

[31] S. Idicula-Thomas, A. J. Kulkarni, B. D. Kulkarni, V. K. Jayaraman and P. V. Balaji, "A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on over expression in Escherichia coli", *Bioinformatics*, 2006, pp. 278-84.

[32] S. Tsantis, D. Cavouras, I. Kalatzis, N. Piliouras, N. Dimitropoulos and G. Nikiforidis, "Development of a support vector machine-based image analysis system for assessing the thyroid nodule malignancy risk on ultrasound", *Ultrasound Med. Biol.*, 2005, pp. 1451-9.

[33] P. M. Kasson, J. B. Huppa, M. M. Davis and A. T. Brunger, "A hybrid machine-learning approach for segmentation of protein localization data", *Bioinformatics*, 2005, pp. 3778-86.

[34] M. E. Mavroforakis, H. V. Georgiou, N. Dimitropoulos, D. Cavouras and S. Theodoridis, "Mammographic masses characterization based on localized texture and dataset fractal analysis using linear, neural and support vector machine classifiers", *Artif. Intell. Med.*, 2006, pp. 145-62.

[35] N Cristianini, and J. Shawe-Taylor, "An introduction to support vector machines and other kernel based learning methods", *Cambridge University Press*, New York, 2000.

[36] Hui-Ling Chen, "A Three-Stage Expert System Based on Support Vector Machines for Thyroid

Disease Diagnosis", *J Med Syst*, 2012, pp. 1953-1963

[37] Y. Shi and R. A. Eberhart, "A modified particle swarm optimizer, in: Evolutionary Computation", *Proceedings. IEEE World Congress on Computational Intelligence*, 1998, pp. 69-73.

[38] A. Ratnaweera, S. Halgamuge and H. Watson, "Self-organizing hierarchical particle swarm optimizer with time-varying acceleration coefficients" *IEEE Trans. Evol. Comput.*, 2004, pp. 240-255.

[39] Y. Shi and R. C. Eberhart,"Empirical study of particle swarm optimization", *Congress on evolutionary computation, Washington D.C.*, USA, 1999, pp. 1945-1949.

[40] K. A. Ron, "A study of cross-validation and bootstrap for accuracy estimation and model selection", *Proceedings of the 14th international joint conference on Artificial intelligence*, 1995.

[41] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification", *Technical report*, National Taiwan University, Taipei, 2003. http://www.csie.ntu.edu.tw/ cjlin/libsvm/.

[42] Christos Bellos, "Categorization of Patients' Health Status in COPD Disease using a Wearable Platform and Random Forests Methodology", *Proceedings of the IEEE-EMBS International Conference on Biomedical and Health Informatics,* China, Jan 2012.

[43] A. Hui-Ling Chen, "A Three-Stage Expert System Based on Support Vector Machines for Thyroid Disease Diagnosis", *J Med Syst*, 2012, pp. 1953-1963.

[44] C. W. Hsu, C. C. Chang and C. J. Lin, "A practical guide tosupport vector classification" *Technical report, National Taiwan University, Taipei*, 2003, Available at http://www.csie.ntu.edu.tw/cjlin/libsvm/.

[45] Chris O Phillips, "Machine learning methods on exhaled volatile organic compounds for distinguishing COPD patients from healthy controls", *Journal of Breath Research*, 2012**,** pp.1-10.

[46] E. Blanca, "Prediction of Chronic Obstructive Pulmonary Disease (COPD) in Asthma Patients Using Electronic Medical Records", *Journal of the American Medical Informatics Association*, 2009, pp.371-379.