CHAPTER 1
INTRODUCTION

1.1 INTRODUCTION

Respiration involves movement of the chest wall to produce a pressure gradient that permits flow and movement of gas. This can be accomplished by the respiratory muscles, through negative pressure ventilation and positive pressure ventilation. Analysis of respiratory function is always performed for the assessment of pulmonary diseases (Grinnan et al 2005).

Pulmonary Function Tests (PFT) are performed in order to diagnose and classify disease processes that impair lung function (Patauchos et al 2008). These tests detect the presence of pulmonary functional abnormalities, quantify their degree and follow the time course of disease. These tests are also helpful for assessing the risk of therapeutic or diagnostic interventions and to monitor the effects of therapy (Anogeianaki 2007). PFT are used to identify the pattern and severity of a physiologic abnormality, but has to be used with other tests to distinguish between the potential causes of the abnormalities.

Flow-volume spirometry is the most common method used to perform pulmonary function tests and is the recommended investigation for diagnosis and categorization of severity of airflow limitation (Feyrouz et al 2003). It is a standardized technique with elaborate guidelines on procedures, evaluation and interpretation. A Spirometer measures the amount of air
moving in and out of the lungs during forced breathing maneuvers (Wagner et al. 2006 and Derom et al. 2008). It allows one to determine how much air can be inhaled and exhaled by generating a flow-volume curve that represents tidal, inspiratory and expiratory phases of breathing. Each FVC manoeuvre requires maximal effort during this "unnatural" breathing manoeuvre which includes maximal inhalation; 2) maximal exhalation for at least one second (for FEV₁); and then 3) continued exhalation for several seconds (for FVC). The shape of the flow volume curves indicates normal or abnormal cases (Arora and Raghu 1996). Respiratory diseases and their severity are interpreted based on the pattern generated by the spirometer along with the prior patient history, physical examination and ancillary diagnostic tests.

The most important parameters of spirometry are Forced Expiratory Volume in 1 second (FEV₁), Forced Expiratory Volume in 6 seconds (FEV₆), Forced Vital Capacity (FVC), Peak Expiratory Flow (PEF) and FEV₁/FVC ratio. FVC is the volume delivered during an expiration made as forcefully and completely as possible starting from full inspiration and is vital for the diagnosis of respiratory abnormalities (Aaron et al. 1999). Normal lungs can empty more than 80% of their volume in 6 seconds. FEV₁ and FEV₆ are the volume of air that can be forcefully expired during the first and the sixth seconds of the maximal expiration respectively (Miller et al. 2005).

Serial measurements of FEV₁ are used to monitor the progression of diseases. But submaximal inhalation effort during the first phase reduces both the FEV₁ and the FVC. A submaximal exhalation blast during the second phase affects the FEV₁ and the subject cannot be forced to perform for repeated trials which might lead to fatigue and breathing difficulties. Also there is a pronounced FEV₁ decline in occupational health hazard cases. In such cases FEV₁ parameter prediction has its own significance. Studies designed to predict FEV₁ help in defining the risk of
pulmonary complication (Baltopoulos et al 2000). The American Thoracic Society has developed a scale to rate the severity of disease based on FEV$_1$ and total lung capacity (Steltner et al 2000). The PFT maneuver is a test that requires effort and cooperation from the patient. For the test to meet the acceptability and reproducibility criteria, patients must empty their lungs completely in at least three maneuvers. Each of these maneuvers can last up to 20 sec and are exhaustive for older and patients with advanced lung diseases.

It has been demonstrated in the literature that FEV$_6$ is an acceptable surrogate for FVC in the diagnosis of restriction and airway obstruction in adults. It has been shown that FEV$_6$ and FEV$_1$/FEV$_6$ may replace FVC and FEV$_1$/FVC respectively in interpreting spirometry. This has the advantage of greater convenience for the patient during the performance of the test, since it would last only 6 seconds (Swanney et al 2000, Hankinson et al 2003 and Soares et al 2008). It has been proved that there is a strong correlation between FEV$_1$/FVC and FEV$_1$/FEV$_6$ ratios (Lundgren et al 2007).

A Spirometer displays lung volumes and flow rates during forced breathing. An error of method occurs in the measurement of these flow rates and volumes when the effect of spirometric transducer resistance influences the respiratory process (Juroszek 2005). During the pulmonary function tests, blocking of little ducts exists in some diseases. During spirometric measurements of air flowing through a pipe-line, the respiratory system is connected serially to the transducer. As a result, the air flows from unblocked alveoli making larger amount of air to be expired. Thus, the transducer flow resistance cooperates with the respiratory resistance in the forced expiration. Therefore all the respiratory parameters are changed depending on the relation between flow resistances of the measuring object and the transducer resulting in the measurement error. The permissible range of transducer resistance is in the range of 2–10% of the overall resistance of the respiratory tract (Juroszek
The error factor in FEV\textsubscript{1} explains the influence of transducer resistance in normal and abnormal subjects. Analysis of error factors due to transducer resistance provides useful information about the interrelationship and significance of spirometric parameters.

Since spirometry is the most widely used screening test to investigate the pulmonary function abnormalities, there is a requirement for the analysis of large database. The spirometric data would also have missing values and patients with lung abnormalities might not be able to repeat the test (Ulmer 2003). Hence there is a need for prediction of unavailable parameters for better diagnosis.

Artificial neural networks (ANN) have been used successfully in prediction and classification of signals, images and data. They are mathematical algorithms that approach the functionality of small neural clusters. ANN is trained from the input parameters and the trained network can be employed for prediction and classification of a set of information. The advantage of neural networks is that they can be used to predict one or more output types through a flexible network of weights, transfer functions and input variables (Sachin et al 2007). They have been used in a great number of medical diagnostic decision support systems (Benardos and Vosniakos 2007). ANN are found to be efficient in the classification of non-periodic and nonlinear types of signals and are extensively used in cardiology, gastroenterology, pulmonology, oncology, neurology, ophthalmology, radiology and in respiratory measurements (Gaetano et al 2001).

A Radial Basis Function (RBF) network is a neural network approached by viewing the issue as a curve-fitting problem in a high dimensional space. Learning is equivalent to finding a multidimensional function that provides a best fit to the training data, with the criterion for “best
fit” being measured in some statistical sense. RBF networks models have proved to be universal function approximators (Rrzvyan et al 2008). The number of hidden nodes, centers and widths of the basis functions used by nodes of RBF networks are tunable. The centers of the radial basis functions can be selected by k- means clustering algorithm, fuzzy c-means clustering algorithm and Self Organizing Map (SOM) algorithm. The k-means clustering algorithm is based on the fact that the data set is partitioned into clusters and the optimal placement of a center is at the centroid of the associated cluster. The k-means algorithm assigns each data point to the cluster whose center is nearest. The fuzzy c-means algorithm partitions the data set into fuzzy clusters, and finds a center by minimizing an objective function. In fuzzy clustering, the data points could belong to more than one cluster, and associated with each of the points are membership grades that indicate the degree to which the data points belong to the different clusters (Marc 2007 and David et al 2007).

Self Organizing Map (SOM) is a clustering algorithm in which different neighborhoods are updated every time a new vector is presented. The SOM is an unsupervised neural network that projects a high dimensional input space onto a two-dimensional output space. The projection is topological preserving, that is, where patterns that are similar in terms of the input space are mapped to geographically close locations in the output space. The SOM is composed of a set of nodes arranged in a geometric pattern, typically two dimensional lattices. Each node is associated with a weight vector W with the same dimension as the input space. Learning in SOM is based on competitive learning where these output nodes of the network compete among themselves to be activated or fired (Kohenen 2000).

Support Vector Machines (SVM) is a machine learning technique based on statistical theory proposed by Vapnik in 1995. It has received
considerable attention in the recent past and is extensively used for the problems of regression and classification. The merit of SVM lies in the theory of the Structural Risk Minimization (SRM) principle in estimating a function by minimizing an upper bound of the generalization error. SRM is an inductive principle for model selection used for learning from finite training data sets. It describes a general model of capacity control and provides a tradeoff between hypothesis space complexity and the empirical error in the training data. Another key property of SVM is that training SVM is equivalent to solving a linearly constrained quadratic programming problem so that the solution of SVM is always unique and globally optimum (Andrea et al 2004 and Chen et al 2008).

The principle of SVM is to find a maximum margin hyperplane for classification. If it is not possible in the given space, the instances are mapped to a higher dimensional space using the kernel function. The kernel functions allow working in a higher dimensional space. This reduces computational complexity and connects the input space and the higher dimensional space directly. SVM chooses a maximum soft margin separating hyperplane in this higher dimensional space that separates the training instances by their classes. The classification of a test sample will be determined by a sign function and is defined by the parameters of the hyperplane. The instances closest to the hyperplane are called support vectors and are vital for training (Vapnik 1995 and Schölkopf et al 1997). Support vector regression techniques are most successful in complex and dynamic environments which can accommodate and recover from infrequent predictive errors.

Principal Component Analysis (PCA) is a statistical method used to transform the input space into a new lower dimensional space and has been used to identify and summarize many inter-relationships that exist among individual variables. In this, inter-correlated variables are combined into a
smaller number of new variables called principal components. The first principal component accounts for much of the variability in the data and each succeeding component accounts for the remaining variability. The uncorrelated variables are linear combinations of the original variables and the last of these variables can be removed with minimum loss of real data in order to identify new meaningful underlying variables (Aguado et al 2008).

PCA of high-dimensional data is an ingredient of many signal processing applications and strives to extract the principal directions in the data space where the variance of the data is maximal, thus paving the way for dimension reduction and data compression (Burstyn 2004). PCA pursues an online approach where an estimate of the principal directions is updated after each presentation of a data point. This method is especially suited for high-dimensional data, since the computation of the large covariance matrix can be avoided, and for the tracking of nonstationary data. The principal components correspond to the directions in which the projected observations have the largest variance which are the eigenvectors of the empirical covariance matrix (Tharrault et al 2008).

Since obstructive and restrictive diseases have a high morbidity and mortality early diagnosis and determination of the severity of such diseases are important. Classification of spirometric data into normal and abnormal is essential in the analysis of respiratory mechanics (Sahin et al 2009).

SVM have high classification accuracies and less prediction errors due to the properties like efficient solutions, relatively few adjustable parameters and the interchangeable use of kernel functions, which define a mapping of the data to a higher-dimensional feature space. It discriminates the data by creating boundaries between classes rather than estimating class conditional densities and needs considerably less data to perform accurate
classification. With a suitable kernel, SVM can separate in the feature space the data that was non separable in the original input space. A kernel function has a good performance if the number of support vectors calculated by using the corresponding transformation is few and the classification of the test data is successful. Support vector set is enriched by those training examples that cannot be classified by the model correctly (Akay et al 2009).

In this work, an attempt has been made to analyze the diagnostic relevance of spirometric investigations using support vector regression, principal component analysis and support vector classification.

1.2 OBJECTIVES OF THE THESIS

The following are the objectives of the thesis:

- Record the respiratory functions of normal, obstructive and restrictive lung conditions using flow-volume spirometry,
- Predict Forced Expiratory Volume in one second (FEV$_1$), using support vector regression,
- Predict Forced Expiratory Volume in six seconds (FEV$_6$), using neural networks and support vector regression,
- Analyze the variations in error factor due to the influence of transducer resistance,
- Perform feature selection from the measured parameters using principal component analysis, and
- Classify pulmonary abnormalities using support vector machines with the predicted FEV$_6$ and FEV$_1$ values.
1.3 ORGANIZATION OF THE THESIS

The work reported in the thesis is organized into 5 Chapters: Chapter 1 deals with brief introduction, objectives and organization of the thesis. Chapter 2 discusses a brief review of the literature on pulmonary function test, spirometry, support vector regression, artificial neural networks, principal component analysis and support vector classification. Chapter 3 describes the methods and protocols, explains the theoretical evaluation of data obtained from spirometry and analysis of the data using support vector machines, neural networks and principal component analysis. Chapter 4 focuses on the results obtained through the analysis. The summary and conclusion drawn from the analysis are discussed in chapter 5. Also, the scope for future work is included in this chapter.