Evaluating Spirometric Trends in Cystic Fibrosis Patients

S. Zarei*, M. Abouali*, A. Mirtar†, J. Redfield‡, D. Palmer§ and D. J. Conrad¶ P. Salamon‡,
*Computational Science Research Center, San Diego State University
†Electrical and Computer Engineering Department, University of California San Diego
§Department of Biology , San Diego State University
¶Division of Pulmonary and Critical Care Medicine, Univeristy of California San Diego
‡Department of Mathematics and Statistics , San Diego State University

***Abstract*—*In this research, we applied both supervised and unsupervised machine learning methodologies to spirometric data from patients with cystic fibrosis (CF). We developed an ensemble of neural networks to evaluate the severity of chronic CF within an individual, given the appropriate clinical input data, and a series of reference equations to describe the CF patient's pulmonary function at different ages, heights, and sex groups in order to determine longitudinal spirometric trends. The neural networks were able to be eighty-eight percent accurate when evaluating chronic disease severity and our regression analysis revealed several trends, such as in females with CF, obstruction and functional airflow movement within the lungs generally tends to deteriorate at an accelerated rate compared to males with CF. Our findings have the potential to serve as useful reference tools to physicians in the diagnosis and treatment of cystic fibrosis.*

Keywords: Cystic fibrosis, neural networks, best-fit regression analysis, spirometric trends

I. Introduction

Cystic fibrosis is an inherited chronic disease that affects the lungs, digestive system, and even the circulatory system of CF patients. Most commonly, CF is characterized by both chronic airway inflammation and recurrent infections, typically leading to permanent structural lung changes and a progressive decline in lung function. Spirometric testing, commonly used by pulmonologists to assess pulmonary function, involves taking measurements to quantify the degree of airway obstruction. Oftentimes, CF patients will go through a series of different tests and treatments in hopes of alleviating their chronic disease symptoms; this results in the accumulation of vast quantities of longitudinal spirometric data and makes this research possible. Defined as the volume of air a patient can forcibly exhale in one second, FEV1 is one of the most significant parameters obtained by spirometry, since it identifies both restrictive and obstructive respiratory symptoms. It is also a powerful predictor of increased risk of lung cancer and other obstructive lung diseases [Miller et al. 2005; Pierce 2004; Wagner et al. 2006]. All CF spirometry data was collected at the Adult Cystic Fibrosis Clinic (ACFC) and Pulmonary Function Laboratory (PFL) at the Veteran's Affairs Medical Center in La Jolla CA, and was made available by Dr. Douglas Conrad, M.D., the ACFC and PFL director.

In different research studies, pulmonary measurements have been collected from healthy individuals of both sexes across range of ages. They provide reference equations for the different aspects of the spirometry test, including FEV1 measurements for males and females of various race/ethnic groups [Hankinson et al. 1999]. On the other hand, CF investigators do not have sufficient statistical analyses for assessing the relationship between pulmonary function outcomes and predictor parameters of interest [Edwards 2000]. The aim of this paper is to discuss some of the important features of statistical analysis on the CF patient FEV1 database. This will include grouping CF patients based on their longitudinal FEV1 data through the use of a clustering method and an evaluation of a regression analysis on the FEV1 data. Next, we find the corresponding reference equations that can be used in predicting the CF patients FEV1 value. Furthermore, we present an ensemble of artificial neural networks to predict the severity of chronic cystic fibrosis within an individual by comparing against fifty patients ranked ordinally by increasing disease severity.

II. Spirometry Regression and Clustering Analysis

The two most important values from a spirometric test are FVC and FEV1. The forced vital capacity (FVC) indicates the maximum volume of air that can be forcibly expired from the lungs. FEV1 represents the forced expiratory volume in the first second. According to the research that was conducted by Hankinson and Odencrantz the best fit regression equation describing the lung function parameter FEV1 was in terms of age and height for different age groups in both males and females of different ethnicities. The general form of their regression equation for Caucasians is as follows:

$$FEV1 = b_0 + b_1 \cdot Age + b_2 \cdot Age^2 + b_3 \cdot Height^2 \quad (1)$$

For research purposes, we use only the corresponding regression equation for Caucasians. This is due to the fact that of all ethnic groups, they hold the highest inherited risk for CF, where approximately 1 in every 25 Caucasians is a carrier for this recessive condition, and 1 in 2,500 are clinically affected [Tsui et al. 1997]. Table I illustrates the corresponding regression equations among healthy Caucasian individuals. In this study, we used the 2004-2009 spirometry test results of patients from the University of California San Diego Adult Cystic Fibrosis Center (UCSD-ACFC), which

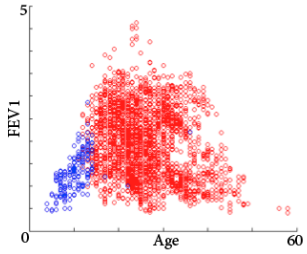

Fig. 1. Age vs. FEV1 with 2 clusters.

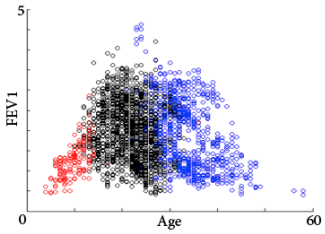

Fig. 2. Age vs. FEV1 with 3 clusters.

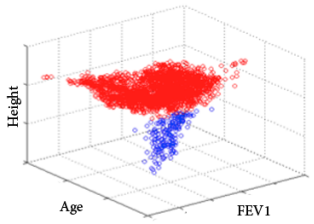

Fig. 3. Age vs. Height vs. FEV1 with 2 clusters.

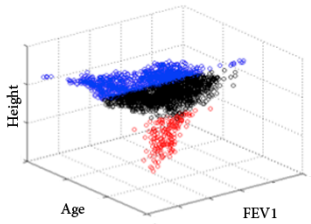

Fig. 4. Age vs. Height vs. FEV1 with 3 clusters.

TABLE I
FEV1 REGRESSION EQUATIONS FOR HEALTHY INDIVIDUALS

Sex	Caucasian $<$ 20 years of age	Caucasian $>$ 20 years of age
Female	$Fev_1 = -0.8710 + 0.06537\times$ $Age + 0.00011496 \times Height^2$	$Fev_1 = 0.4333 - 0.00361\times$ $Age - 0.000194 \times Age^2$ $+ 0.00011496 \times Height^2$
Male	$Fev_1 = -0.7453 - 0.04106\times$ $Age + 0.0004477 \times Age^2$ $+ 0.00014098 \times Height^2$	$Fev_1 = 0.5536 - 0.01303\times$ $Age - 0.000172 \times Age^2$ $+ 0.00014098 \times Height^2$

includes approximately a total of 6,000 samples. Our first attempt was to find the possible clusters within our samples to help us group CF patients based on their lung function. Figures 1 through 4 are the cluster plots using the K-mean method. In data mining, K-means clustering is an algorithm for partitioning (or clustering) N data points into K disjoint subsets S_j containing N_j data points in a way that minimizes the following sum-of-squares criterion.

$$J = \sum_{j=1}^{K} \sum_{n \in S_j} |x_n - \mu_j|^2 \tag{2}$$

where x_n is a vector representing the nth data points, and μ_j is the geometric centroid of the data points in S_j[Bishop1995]. This will result in K clusters in which each observation belongs to the cluster with the nearest mean. As depicted in Figure (1) , the FEV1 values of CF patients can be divided into two different groups one above and one below age 15. On the other hand, Figure (2) shows that when we used three clusters the additional group formed between ages 30 to 60. According to the CF foundation, more than 45% of the CF patient population is age 18 or older, additionally the predicted median age of survival for a person with CF is 37 years [Cystic Fibrosis Foundation]. Therefore, we can refer to the blue (darker) part of Figure (2) as the survival group. In Figures (3) and (4) we plot age vs height vs FEV1 values. As both figures show, age 25 is the main separation line regardless of the number of clusters. This could be due to the growth in height during that age span. In Figure (4), even though we have three clusters, two have similar heights and the one showing age 25 and under falls into a separate cluster, which is easily seen in Figure (3) with only two clusters. By clustering our CF FEV1 data we realized that as the height of the CF patients increases throughout their developing years, between the ages of 0-25, their FEV1 values increase.

III. REGRESSION ANALYSIS OF LONGITUDINAL CF FEV1

In this section, we find the best fit equation for different sex and age groups. According to Figure (1) and (2), the cut-off age should be 15, yet due to the fact that we are trying to compare lung function between healthy individuals and CF patients, we preferred to use the same age grouping as Hankinson and Odencrantz [1999]; that is before and after age 20 which is close to our cut-off age. For each group we conducted a regression analysis to identify the best-fit equation among the following four equations:

Reg 1) $FEV1 = b_0 + b_1 \cdot Age + b_2 \cdot Height$

Reg 2) $FEV1 = b_0 + b_1 \cdot Age + b_2 \cdot Age^2 + b_3 \cdot Height$

Reg 3) $FEV1 = b_0 + b_1 \cdot Age + b_2 \cdot Age^2 + b_3 \cdot Height^2$

Reg 4) $FEV1 = b_0 + b_1 \cdot Age + b_2 \cdot Age^2 + b_3 \cdot Height + b_4 \cdot Height^2$

In order to find the best fit equation for each group, we considered corresponding residual plots, normal plot of residuals, coefficient of determination as well as Akaike's information criterion values. Therefore, we selected a model with the following characteristics:

Lowest SS_{err} (Sum of Square Error) value:

$$SS_{err} = \sum_{i=1}^{n} \epsilon_i^2 \tag{3}$$

Highest coefficient of determination:

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}} \quad \text{where} \quad SS_{tot} = \sum_i (\bar{y} - y)^2 \tag{4}$$

Lowest Akaike information criterion (AIC):

$$AIC = 2K + n \cdot [\ln(2 \cdot \pi \cdot \frac{RSS}{n}) + 1] \tag{5}$$

where k is the number of parameter and n is the number of sample. Generally, having more parameters in a regression equation will result in a higher R^2, and lower RSS. However, the optimal model is one consisting of only necessary parameters. Traditionally we would add a parameter to our model only if it increases the R^2 value by a minimum of 5%. Given a data set, several competing models may be ranked according to their AIC, with the one having the lowest AIC being considered to be the best. The AIC methodology attempts to find the model that best explains the data with the minimum number of free parameters. Table II displays

our regression results for females over the age of 20. Based

TABLE II
FEV1 REGRESSION REGRESSION VALUES FOR FEMALES OVER AGE OF 20

Model	RSS	R^2	AIC
Reg1	0.9318	0.88591	-16.6689
Reg2	0.9169	0.88774	-15.1704
Reg3	0.9137	0.88813	-15.2779
Reg4	0.6325	0.92257	-24.6821

on Table II, residual plots and normal plots show the best fitted equation for females above the age of 20 is $FEV1 = b0 + b_1 \cdot Age + b_2 \cdot Age^2 + b_3 \cdot Height + b_4 \cdot Height^2$, since it captures the lowest AIC value, highest regression coefficient and lowest RSS values. The same model was selected as the best fit equations for females under the age of 20 as well.

$$\text{Female} > 20 :$$
$$FEV1 = 664.0178 - 0.0232 \cdot age - 0.0010 \cdot age^2 - 8.1568 \cdot Height + 0.0251 \cdot Height^2 \quad (6)$$

$$\text{Female} < 20 :$$
$$FEV1 = 0.00008 + 0.49077 \cdot age - 0.1219 \cdot age^2 + 0.02319 \cdot Height + 0.00004 \cdot Height^2 \quad (7)$$

We repeated the same regression analysis on the male sample data. Using the Table III regression results and their corresponding residual normal plots, we selected the second regression model as the best to define the FEV1 for male CF patients over, as well as under the age of 20.

TABLE III
FEV1 REGRESSION VALUES FOR MALES OVER AGE OF 20

Model	RSS	R^2	AIC
Reg1	1.5006	0.75899	-13.8053
Reg2	1.4055	0.77427	-14.4240
Reg3	1.4057	0.77423	-14.4176
Reg4	1.4029	0.77468	-12.4970

$$\text{Male} > 20 :$$
$$FEV1 = 9.6464 - 0.0634 \cdot age + 0.0004 \cdot age^2 - 0.0332 \cdot Height \quad (8)$$

$$\text{Male} < 20 :$$
$$FEV1 = -0.119 - 0.0339 \cdot age + 0.0054 \cdot age^2 - 0.0094 \cdot Height \quad (9)$$

IV. HEALTHY VS. CF FEV1 COMPARISONS

After finding the best fit equation of FEV1 for Cystic Fibrosis patients, using the same ages and heights, we found the FEV1 value for healthy individuals by using the lung function parameter equation found by Hankinson and Odencrantz [1999] . We then compared the ratio of the two regression equations. As Figures (5) and (6) depict, the FEV1 values for healthy individuals are always higher than the FEV1 of those with Cystic Fibrosis. As shown in Figure (7), average females with CF begin with almost 75 percent lung function at age 8

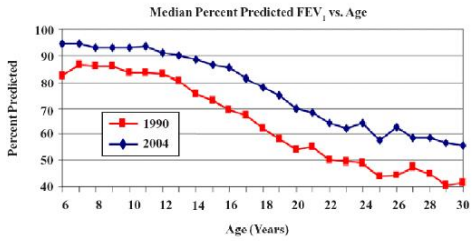

Fig. 9. Median percent predicted FEV1 vs. Age for years 1990 and 2004.

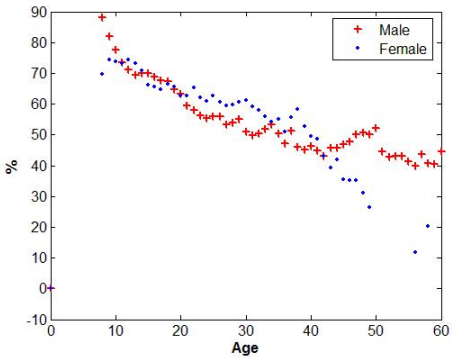

Fig. 10. Median percent predicted FEV1 vs. Age for year 2008.

and that number drops down to only 10 percent by age 60 (for the survival group). On the other hand, Figure (8) shows FEV1 values for men with CF begin with almost 90 percent lung function at age 8 and only 40 percent of lung function by age 60. Figures (9) and (10) represent the functional lung volume of CF patients as the percents of the normal lung for the years 1990, 2004 and 2008 respectively. We can see a significant improvement compared to year 1990 when at age 30, the average CF patient had only 40 percent normal lung function compared to 2004 and 2008 where this value has increased to 55 percent. This may be due to the advanced treatments developed since 1990 that help CF patients to control the progress of the disease.

V. ARTIFICIAL NEURAL NETWORKS

In this section we develop an ensemble of artificial neural networks (ANNs) to predict chronic disease severity within cystic fibrosis patients as an experienced pulmonary physician would. ANNs are a form of machine learning algorithm based on the functionality and structure of a biological neural network, as observed in the brain. Used to observe complex trends and patterns in a set of data, they are capable of applying sets of non-linear equations to inputs to achieve a desired outcome. These equations can be used to apply to further data. The data was collected from the Adult Cystic Fibrosis Clinic (ACFC) and Pulmonary Function Laboratory (PFL) at the Veteran's Affairs Medical Center. Fifty patients were selected and ordinally ranked by Dr. Douglas J. Conrad, director at the ACFC and PFL, in order of increasing disease severity ranking 1 to 50 as a training dataset for the ANNs. The 50 patients and their corresponding 14 variables were compiled as a matrix, along with their actual rankings, and

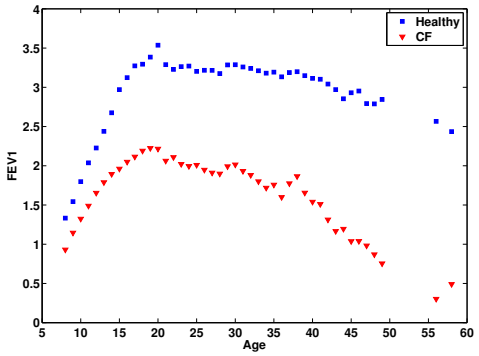

Fig. 5. CF FEV1 compare to healthy FEV1 (female).

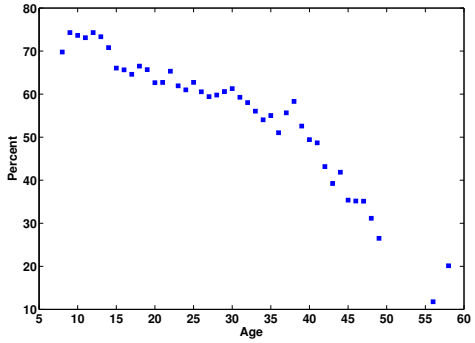

Fig. 6. Functional lung volume of CF patients as a percent of the healthy lung (female).

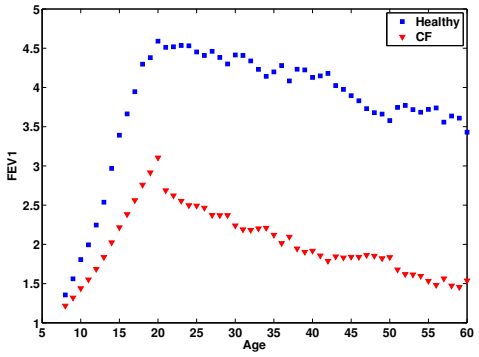

Fig. 7. CF FEV1 compare to healthy FEV1 (male).

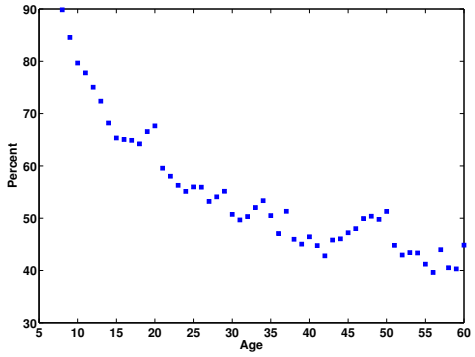

Fig. 8. Functional lung volume of CF patients as a percent of the healthy lung (male).

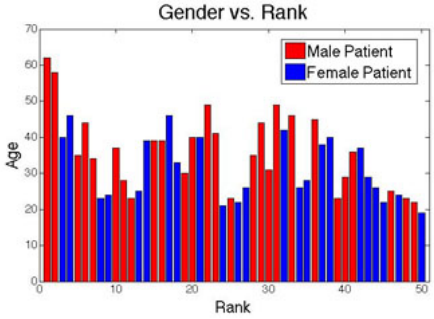

Fig. 11. Age vs. Rank vs. Gender for the 50 patient data set

imported to Matlab. Such variables included results from lung function tests (FEV1, FVC, FEV1/FVC), physical descriptions (age, height, weight, gender, BMI) and longitudinal regression values based on FEV1 vs. time graphs (m, b, r^2, se_m, se_b). For each patient, only the best FEV1 value from the previous year was considered.

VI. ANN Training

To obtain the artificial neural networks, a progression of training, validating and testing steps were taken to develop the ability to predict with an acceptable amount of error. In training, each network is supplied with a set of data as inputs, and through a series of equations, returns an answer. Once the tested outputs of the network accurately reflect the answers provided in the training data, the ANNs can then be used to classify future data. For an ANN to follow the trends in cystic fibrosis data, the variables were run through a series of equations, deemed "layers." Four matrices of random numbers were generated, two representing weights and two being biases. Let w_1 and w_2 denote the weight matrices, b_1 and b_2 the biases, and "in" represents the vector of one patient's inputs. The layout of a single-hidden layer ANN is as follows:

$$\text{hidden layer} = squash((in \times w_1) - b_1) \tag{10}$$

$$\text{CF severity} = (\text{hidden layer}) \times w_2 - b_2 \tag{11}$$

$$squash(x) = \frac{1}{1 + e^{-x}} \tag{12}$$

Once the severity has been run through the above layers for each patient, the ANN returns its predicted severity, and the error is calculated between its prediction and the actual answer. The weights and biases are adjusted using Matlab's fminsearch optimization function, and after each adjustment, the squared error is recalculated between the ANN outputs and the actual provided patient severity. Fminsearch is set to repeat these adjustments until a minimum in the total calculated error is found. However, if the entire set of 50 patients were to be used in training a network, there would be no unknowns upon which

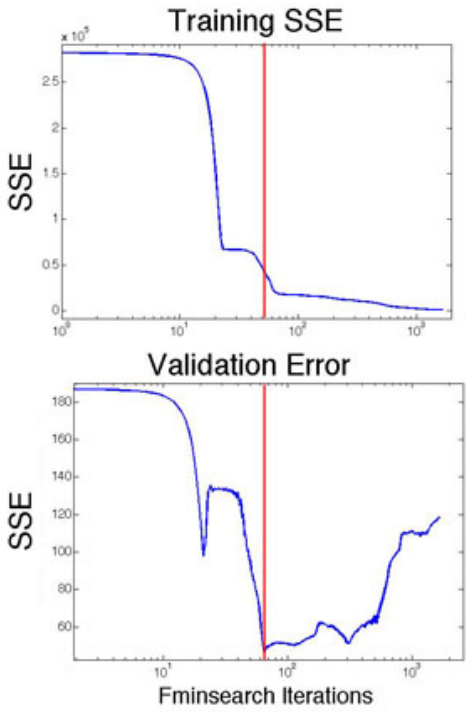

Fig. 12. Observed SSE during Training and Validation

to test its accuracy. For this reason, only thirty of the fifty patients, or 3/5 of the original data set, were randomly selected and used to train each ANN. The remaining 20 patients are randomized and split evenly into validation and testing groups. In validation, the purpose is to halt the fminsearch function once a network begins to over-train. Between the fminsearch iterations of the training set, the squared error is calculated and recorded for the ten validation patients. Once the weights and biases have become overly specific for the training set, the validation error will increase and halt the network training, as shown in Figure (12). The test set for the network is recorded along with the adjusted parameters.

VII. Initial ANN Testing

A set of twenty networks' parameters and test sets was compiled. Each patient defined to be in a test set was run through the layers using the corresponding network's weights and biases, and averaged with the other participating networks. Thus, each ANN only "voted" on the severity of inputs that were not used in its training or validation processes. The averaged output consists of fifty patients, to be compared against the actual severities provided on the original fifty-patient data sheet (Figure (14)).

VIII. Subsequent Inputs

Following the testing of the original networks, the importance of the 14 training variables was determined. Using the computational program R and the randomForest toolbox, FVC, FEV1, and obstruction ratio were found to hold the strongest predictive power for CF severity (Figure (13)). A new ensemble of 50 ANNs were trained and tested using only the FVC, FEV1, and obstruction ratio as training features

First Inputs	Node Purity	Second Inputs	Node Purity
FVC	3018.8	Multiproduct	2308.5
FEV1	1960.8	FEV1	1217.0
Obstruction Ratio	1086.6	FVC	960.5
Age	616.3	PowerProduct	832.5
BMI	593.3	Obstruction Ratio	622.3
Weight	452.0	Brasfield	563.9
m	381.0	Cystic	481.3
r^2	377.9	Age	456.1
b	299.5	BMI	293.1
Patient ID	290.1	Overall	254.1
Height	258.6	Patient ID	160.1
sey	219.2	Height	145.4
seb	201.0	Linear	131.4
sem	155.7	Exp	73.1
Gender	21.5	Gender	16.5
		LgLS	13.2

Fig. 13. Inputs used in both matrices

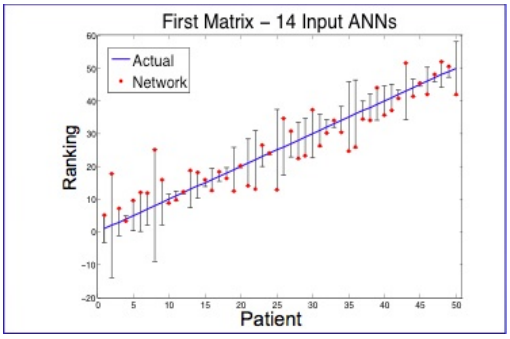

Fig. 14. Severity prediction from ANNS

(see Figure 15). A new dataset was then assembled including several new inputs. Multiproduct and powerproduct attempt to place emphasis on age, and were generated from the equations:

$$\text{multiproduct} = Age \times FEV1\% \tag{13}$$

$$\text{powerproduct} = FEV1\% \times e^{\frac{Age}{10}} \tag{14}$$

Other variables included the Brasfield score and its components. The randomForest toolbox predicted multiproduct, FEV1, FVC, powerproduct, obstruction ratio, and the overall Brasfield score as the most important variables of the new set (Figure (13)). An ensemble of ANNs were trained from these six variables and tested for their performance (Figure (16)). For each of the 3 sets of inputs, the R^2 values and ranking accuracies were calculated, shown in Table IV. The ranking accuracy is defined by the ability of the ANNs to identify the more severe case of CF for any two patients.

TABLE IV
Results of ANN Voting

Dataset	Inputs	Train Time (min)	R^2	Ranking Accuracy %
1	14	360	0.8261	86.67
1	3	10	0.7845	85.14
2	6	30	0.8109	88.48

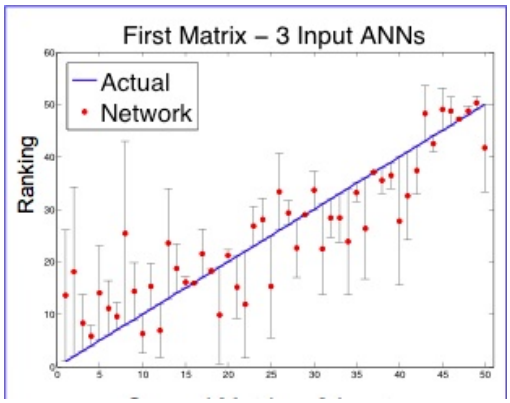

Fig. 15. Severity prediction from ANNS

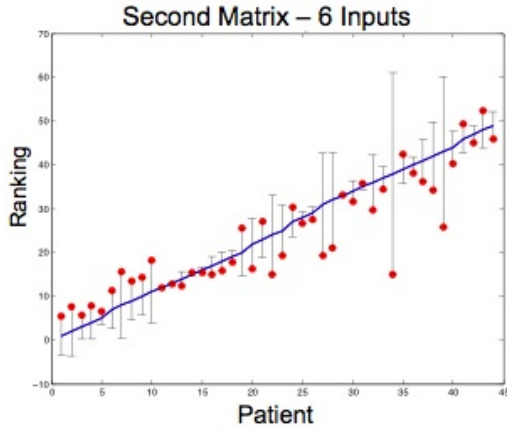

Fig. 16. Severity prediction from ANNS

IX. Conclusion

In conclusion, we were able to find the best fit equations for identifying lung function parameters (FEV1) for both male and female among CF patients. Using these equations we were able to see the overall trend of reduction in their lung function as a over time. Females' FEV1 and FVC values decline faster than males when afflicted with CF. The overall trends of CF lung function have improved due to advanced treatments discovered in more recent years. Using our reference equations, clinicians can predict CF patients' FEV1 and use it as a reference tool for evaluating their treatments. Ensembles of neural networks were able to be trained from the provided inputs and accurately vote upon unseen CF patients. The variables FEV1, FVC, and obstruction ratio appeared to hold the greatest ability to train the ANNs from the original list of inputs. Of the second list of inputs, the Brasfield Index, multiproduct, and powerproduct were also found to be useful in ANN training. The patient data provided has shown potential leeway for training ANNs to perform other medical analysis, such as predicting CF exacerbations per year or the severity of a given patient in five years. These ANNs can be programmed into a GUI available for practitioner use.

Acknowledgement

The author wishes to acknowledge the helpful comments and suggestions made by Rojeen Zarei. This work was made possible by NSF grant UBM 0827278 to A. M. Segall and P. Salamon. We would like to also thank the SDSU UBM and Cystic Fibrosis Groups for their guidance and many helpful discussions.

References

[1] Bishop, C. M. Neural Networks for Pattern Recognition. Oxford, England: Oxford University Press, 1995.

[2] Cystic Fibrosis Foundation. [Online] [Cited: May 27, 2012.] http://www.cff.org.

[3] Edwards, L. J. Modern statistical techniques for the analysis of longitudinal data in biomedical research 30: 330-344, 2000.

[4] Eigen, H., H. Bieler and D. Grant, Spirometric pulmonary function in healthy preschool children. Am J Respir Crit Care Med 163: 619-623, 2001.

[5] Glindmeyer, H. W., J. J. Lefante, C. McColloster, R. N. Jones, H. Weill. Blue-collar normative spirometric values for Caucasian and African-American men and women aged 18 to 65. Am J Respir Crit Care Med 151: 412-422, 1995.

[6] Glenny, R. W., S. L. Bernard, and H. T. Robertson. Pulmonary blood flow remains fractal down to the level of gas exchange. J Applied Physiology 89: 742-748, 2000.

[7] Hankinson, J. L., J. R. Odencrantzand and K.B. Fedan. Spirometric Reference Values from a Sample of the General U.S Population. Am. Jr. Respi. Crit. Care Me 159: 179-187, 1999.

[8] Miller, M. R., J. Hankinson, V. Brusasco, F. Burgos, R. Casaburi, A. Coates, R. Crapo, P. Enright, C.P.M. van der Grinten, P. Gustafsson, R. Jensen, D. C. Johnson, N. MacIntyre, R. McKay, D. Navajas, O.F. Pedersen, R. Pellegrino, G. Viegi, J. Wanger. Standardisation of spirometry. European Respiratory Journal 26: 319-338, 2005.

[9] Nelson, S. B., R. M. Gardner, R. O. Crapo and R. L. Jensen. Performance evaluation of contemporary spirometers. Chest 97: 288-297, 1990.

[10] O'Donnell, D. E., M. Lam, K. A. Webb, M. Lam and K. A. Webb. Spirometric correlates of improvement in exercise performance after cholinergic therapy in chronic obstructive pulmonary disease. Am J Respir Crit Care Med 160: 524-549, 1999.

[11] Pierce, R. Spirometer: An essential clinical measurement. Australian Family Phyician: 34, 535-539, 2004.

[12] Smith, J. J., S.M. Travis, E.P. Greenberg and M.J. Welsh. Cystic fibrosis airway epithelia fail to kill bacteria because of abnormal airway surface fluid. Cell 85: 229-236, 1996

[13] Tsui, L. C., P. Durie. Genotype and phenotype in cystic fibrosis, Hosp Prac 32: 115-142, 1997.

[14] Wagner, N. L., W. S. Beckett and Steinberg, R. Using Spirometry results in occupational medicine and research: common errors and good practice in statistical analysis and reporting. Indian Journal of Occupational Environmental Medicine 10: 5-10, 2006.