GOLD CLASSIFICATION OF COPDGENE COHORT BASED ON DEEP LEARNING

Jun Ying[1,2], Joyita Dutta[1], Ning Guo[1], Lei Xia[2], Arkadiusz Sitek[1], Quanzheng Li[1]
Quanzheng Li: Li.Quanzheng@mgh.harvard.edu
[1] Nuclear Medicine and Molecular Imaging Radiology Department, Massachusetts General Hospital, Boston, MA, USA
[2] Medical Support Center, 301 Hospital, Beijing, China

ABSTRACT

This study aims to employ deep learning for the development of an automatic classifier for the severity of chronic obstructive pulmonary disease (COPD) in patients. A three-layer deep belief network (DBN) with two hidden layers and one visible layer was employed to generate a model for classification, and the model's robustness against exacerbation was analyzed. Subjects from the COPDGene cohort were staged using the GOLD 2011 guidelines. 10,300 subjects with 361 features each were included in the analysis. After feature selection and parameter optimization, the proposed classification method achieved an accuracy of 97.2% by using a 10-fold cross validation experiment. The most sensitive features as revealed by the DBN weights were consistent with the clinical consensus as per previous studies and clinical diagnosis rules. In summary, we demonstrate that the DBN is a competitive tool for exacerbation risk assessment for patients suffering from COPD.

Index Terms—Chronic Obstructive Pulmonary Disease (COPD), Global Initiative for Chronic Obstructive Lung Disease (GOLD), deep learning, Deep Belief Networks (DBNs), classification

1. INTRODUCTION

Chronic Obstructive Pulmonary Disease (COPD) is a kind of inflammatory lung disease characterized by an irreversible blockage of airflow in the lungs. It is also a life threatening lung disease and yet widely under-diagnosed in general [1]. In clinical practice of COPD, precise assessment of severity stage is crucial to diagnosis, prediction of the risk of future exacerbations, and therapeutic guidance. Hence, in the 2010s, the National Institutes for Health and Clinical Excellence (NICE) upgraded COPD guidelines to specifically address the multidimensional assessment of GOLD [2]. The new strategy stipulates that COPD management and treatment should consider both disease impact determined by assessment of symptoms and activity limitation and future risk of exacerbations determined from airflow limitation or exacerbation history.

Recent research has shown that the usage of complementary information from different modalities in COPD classification is efficient [3-5]. Although previous research demonstrated the effectiveness of their methods in multimodal COPD classification, the main limitation is that they considered only simple and low-level features, such as lung function, exhaled volatile organic compounds, health status or demographic factors. In our study, we hypothesize that, inherent in the original clinical features, there exists hidden or latent high-level information related to COPD severity. The inherence can help us set up a more robust model for building a computerized support system for clinical decision making. Deep learning architectures for classification have received great attention in various fields due to their representational power. In this work, we exploit deep learning for feature representation to enhance classification accuracy [6].

A deep learning architecture named Deep Belief Network (DBN) is reported to achieve highly competitive performance [7, 8]. A DBN can be regarded as a highly complex nonlinear feature extractor where each layer of hidden units learns to represent features that capture higher order correlations in the original input data. In this article, we propose to use DBNs for COPD GOLD 2011 classification. Our objectives are to use COPDGene data to:
(1) Evaluate the performance of DBNs in COPD clinical decision making.
(2) Examine the critical features in COPD patients as revealed by deep learning for GOLD classification, and determine their potential relationship with classification for clinical diagnostic practice.

2. MATERIALS AND METHODS

2.1. Data

In this work, we study the Genetic Epidemiology of COPD (COPDGene) cohort, which is funded by National Heart, Lung, and Blood Institute (NHLBI) to investigate the genetic susceptibility of COPD [9]. 10,300 subjects enrolled in research recording were distributed over the full spectrum of disease severity and across both genders. The cohort included 2/3 non-Hispanic white and 1/3 African American individuals. Every study participant had a data record consisting of 361 features including self-administered

questionnaires of demographic data and medical history, questionnaires on symptoms, medical record review, physical examination, and spirometric measures of lung function before and after the administration of a short acting inhaled bronchodilator. (see full data collection forms on COPDGene web site at www.COPDGene.org).

All subjects were labeled using the GOLD 2011 severity classification. This combined assessment of GOLD 2011 groups patients into four categories [10]. Category A represents low symptoms and low risk, category B represents high symptoms and low risk, category C represents low symptoms and high risk, and category D represents high symptoms and high risk. For more detailed classification, the subjects are subdivided into higher risk categories (C & D). Any patients in the C or D categories that meet only the FEV1 criteria are assigned C1 and D1 subcategories respectively. Those that meet exacerbation criteria only are assigned C2 and D2 subcategories. Finally, those meeting both exacerbation and FEV1 criteria, are labeled C3 and D3. Thus, by the GOLD 2011 strategy, subjects can be classified into 8 subtypes (A, B, C1, C2, C3, D1, D2, D3) with different levels of symptoms and risk.

2.2. Feature Selection

For higher performance, we adopt missing data strategies and feature selection. Before feature selection, we apply a strategy for handling missing data. We find the features with missing data covering more than 90 percent of the whole record number and discard them and then fill the missing data of a remaining feature with the mean value of existing data in the same feature. We then normalize the data to generate zero mean and unit variance.

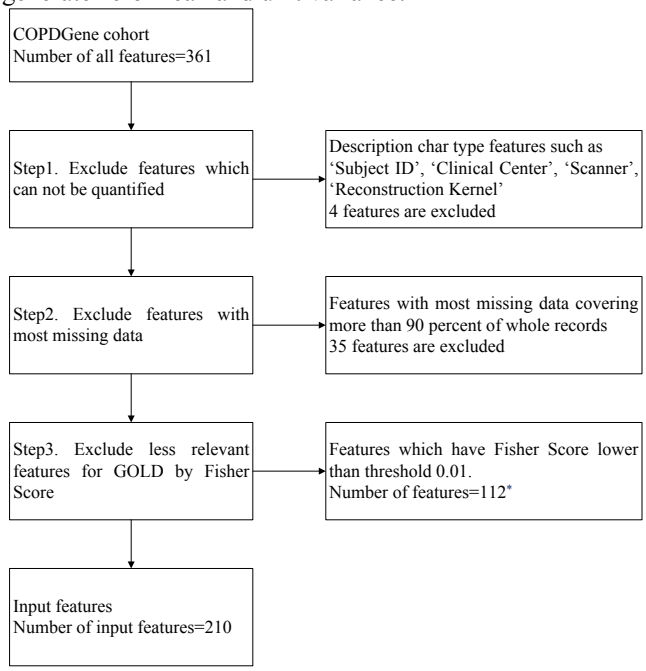

Fig.1. Flow chart describing the feature selection process

The most sensitive features are selected by a discriminative score defined by Fisher Score Criterion [11]. If a feature has a high *F* score, it implies that the data points for this feature have high between-class scatter and low within-class scatter. Features with higher *F* scores tend to be more sensitive and therefore more reliable for classification. In the current study, we regard the high Fisher Score features as the most predominant attributes for the classification of various stages of COPD.

2.3. Deep Belief Networks

Our proposed method performs deep learning by reducing the dimensionality of the input feature vectors using series layers of densely connected DBNs, composed of multiple Restricted Boltzmann Machines (RBMs) [12].

2.3.1. RBM

An RBM is set up by a Markov random field with trainable weights and with nodes separated into a visible layer and a hidden layer [13]. In RBMs, $p(v,h;\theta)$ is the joint distribution between nodes over the visible units v and hidden units h. Given the model parameters θ, $p(v,h;\theta)$ is defined with an energy function $\mathrm{E}(v,h;\theta)$ as follows:

$$p(v,h;\theta) = \frac{exp(-\mathrm{E}(v,h;\theta))}{Z}. \quad (1)$$

For a standard binary-valued type RBM (Bernoulli-Bernoulli), the energy function is given by:

$$\mathrm{E}(v,h;\theta) = -\sum_{i=1}^{V}\sum_{j=1}^{H} w_{ij} v_i h_j - \sum_{i=1}^{V} b_i v_i - \sum_{j=1}^{H} a_j h_j. \quad (2)$$

Z is defined as a normalization factor that can be calculated as:

$$Z = \sum_{i=1}^{V}\sum_{j=1}^{H} exp(-\mathrm{E}(v,h;\theta)), \quad (3)$$

where w_{ij} denote the symmetric interaction term between visible unit v_i and hidden unit h_j, V and H are the numbers of visible and hidden units respectively, b_i and a_j are the bias terms.

2.3.2 DBNs

The multi-layer perceptron for DBNs consist of many layers of RBMs. We construct a three-layer DBN with two layers of hidden units in this study. The input vector is processed layer by layer from the first to final one, and then the output is transformed into a multinomial distribution.

Conventionally we train the baseline system in DBNs by greedy layer-wise unsupervised training [14]. The training procedure is based on a generative method and backpropagation learning. We first train RBM stacks using a generative approach and then fine-tune all the parameters jointly between the true result and the predicted values over every category by maximizing the frame-level cross entropy. To evaluate the performance of our approach, in the training and testing phases, 10 different data randomizations of stratified 10-fold cross-validation (10FCV) were performed.

3. RESULTS

3.1. Classification Performance

The Fisher Score method was employed to select a subset of most informative or discriminative features from the original feature set to improve classification performance. Every features was assigned *F* scores corresponding to

different classifications. In every classification, we ranked all features in descending order of F scores as shown in Fig. 2. The threshold was set to 0.01. Those features with higher F scores were regarded as useful features for classification and were selected as input feature vectors for the DBN.

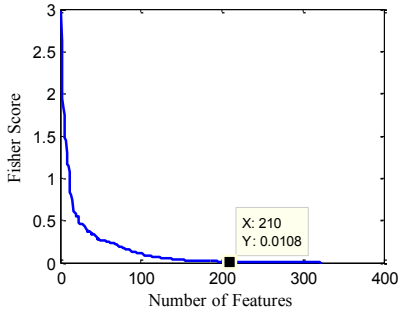

Fig. 2. Fisher Score order for a subset of more important features

Fig. 3 shows the test accuracy of the DBN before and after feature selection for GOLD 2011 classification. Accuracy was calculated under varying numbers of training iterations from 1 to 1000. The accuracy curve remained stable as the number of iterations passed 100. The 10FCV mean classification accuracy improved by 6.1% and maximum classification accuracy by 97.2% upon using the Fisher Score for GOLD 2011.

Fig. 4 shows the accuracy of GOLD 2011 classification with different numbers of training data points. The number of training datapoints was varied from 100 to 9000, while the the test number was held fixed at 1000. In deep learning, the performance of classification is closely linked to the size of the training dataset. For GOLD 2011, when the number of imputed training data exceeded 2000, the accuracy can be increased to more than 90%.

3.2. Ranking of Features

DBNs embed the distribution probabilities as interaction terms between the visible unit and the hidden unit. The units in the first layer weight matrix represent the symmetric interaction between the visible and hidden units and express the distribution probabilities. Accordingly the formula, w_{ij} has a monotonically increasing relationship with the conditional probabilities p. Higher w_{ij} values lead to higher conditional probabilities describing stronger interactions.

To find the most valuable features in the DBNs, we accumulate the weight of each row in the first layer weight matrix as the evaluation index of the importance of each feature. We rank all input features in terms of this index, group the top 50 features into 6 subtypes, and computed the number of features in every subtype. Fig. 5 shows the numbers of features in six subtypes. More than 95% of top 50 valuable features are concentrated in three subtpyes including lung function, respiratory diseases inquiry, and health status measure.

Table I lists the top 10 specific valuable features based on row summation. The top 10 features for GOLD 2011 include physical health inquiry, pulmonary function, clinical symptoms, exacerbation frequency. Features closely associated with SGRQ or other health status inquiry are on the top of valuable features for GOLD 2011.

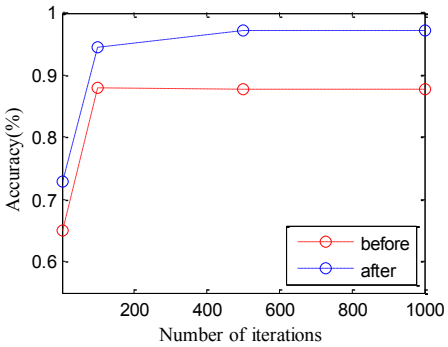

Fig. 3. Accuracy of GOLD classification with different training iterations

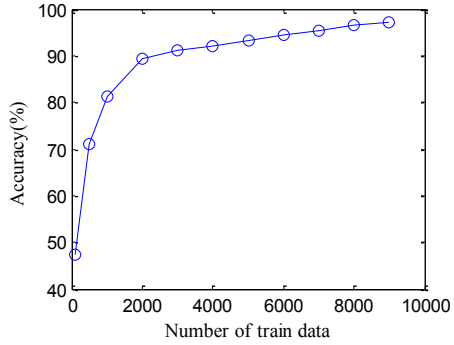

Fig. 4. Accuracy of GOLD 2011 classification with different numbers of training datapoints

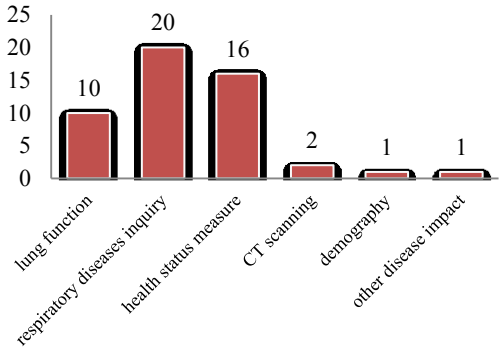

Fig. 5. Numbers of subtypes in the top 50 most important features

TABLE I The top 10 most important features extracted by DBNs

order	Name of features	Description
1	SGRQ_scoreImpact	SGRQ score: Impact
2	SGRQ_scoreActive	SGRQ score: Active
3	MedorTreatAttack	How you ever required medicine/treatment for attacks of wheezing/whistling in chest?
4	ChstWheezyWhist	Ever had wheezing or whistling in chest?
5	LtdUphill	Have shortness of breath

order	Name of features	Description
		when hurrying or slight uphill?
6	CopdAge	At about what age did COPD start?
7	SF36_PCS_score	SF-36 Physical Health Aggregate (PCS) Score
8	FEV1pp_utah	FEV1 % pred, post-Utah
9	Exacerbation_Frequency	Exacerbation Frequency
10	SmokCigNow	Do you now smoke cigarettes (as of one month ago)?

4. DISCUSSION

In this paper we present a new approach to classify different severity stages of COPD based on deep learning. We used supervised feature learning for computational phenotype discovery and classification from noisy, irregular, and high-dimensional multimodal data.

4.1. Feature Selection

In many applications in machine learning and data mining, one is often confronted with very high dimensional data. The philosophy behind feature selection is that not all the features are useful for learning. High dimensionality increases the time and space requirements for processing the data [15]. In this study, more than 300 features of over 10,000 subjects are engaged as deep learning input data. We used the Fisher Score criterion to identify the most valuable features and used these as inputs for DBN training. It was observed that the DBN integrating Fisher Score algorithm can achieved better performance accuracy when compared with only DBN method.

4.2. Classification Performance

We hypothesize that higher-level multivariate features may represent characteristic of COPD and may provide complex data-driven phenotypes representing the disease variants and subtypes. We construct a three layer DBN architecture and produce features that are domain-recognizable subtypes. An accurate mathematical DBN model is trained by running multiple iterations. After 500 iterations of training, the prediction model was completed to ensure that the results had a reliable and reproducible output and the accuracy converged to stable level. A major outcome of this study is that the accuracy of classifications are up to 97.2% for GOLD 2011, which suggests a high level of model accuracy, thereby revealing the potential of deep learning to in mimicking the thinking process of certified professionals who identify diseases and diagnose.

4.3. Top Sensitive Features for GOLD 2011 Classification

In this study, our goal was to identify critical factors related to COPD and its comorbidities. We ranked the top 50 sensitive features for classification in accordance with indices from first layer weights. All of them were found to be strongly linked with COPD classification. In clinical practice, GOLD 2011 is dependent not only on pulmonary function but also on state of an illness and health status [16]. The distribution of subtype feature numbers shows that for GOLD 2011 there are more features based on respiratory disease inquiries and health status measures. This distribution shows that deep learning can find the most sensitive features as well as make a highly accurate classification, which reveals potential correlation information between the features and disease diagnosis.

Furthermore, from the top 10 features for classification, we can find the correlation of specific features and clinical decision making practice. For GOLD 2011, among more than 300 input features, aside from airflow limitation variables related to FEV1, the most sensitive variables were health status survey results, such as SGRQ score, respiratory disease inquiries, and the Short Form Health Survey (SF36) [17]. The SGRQ is the most commonly used in COPD studies to assess differences in health status between the various scenarios [10]. This measure therefore clearly influences the new GOLD classification. Exacerbation frequency is also found as being most sensitive to GOLD 2011 classification [18]. GOLD 2011 uses two treated exacerbations per year as a definition of high risk, assigning these patients into groups C or D. It has been argued that the exacerbation history should also include untreated exacerbations.

The topmost sensitive features found by deep learning are consistent with the classification principle factors in current COPD clinical diagnostics. The strong association of COPD GOLD classification with such sensitive features has been supported by previous studies.

However, COPD is a heterogeneous disease, and to better classify the patients for prognostic purposes and to tailor treatment, the new classification system GOLD 2011 adds new dimensions to the stratification of patients with COPD. The GOLD 2011 assessment system incorporates symptoms and risk of exacerbations, in addition to an assessment of airflow limitation. Airflow limitation in terms of FEV1 in percent predicted decreased with increasing levels of GOLD 2011 severity [19]. The GOLD 2011 update on COPD now bases the classification of the disease on the degree of dyspnea and history of exacerbations in addition to FEV1. Deep learning has been used to determine important factors correlated with COPD classification. The results show that the DBN provides useful information on evaluating contributing features to GOLD classification. Furthermore, our determination of the most sensitive features relative to GOLD classification is consistent with the consensus classification rule in current clinical practice.

ACKNOWLEDGMENTS

This work is supported by NIH R01EB013293. The COPDGene Study is funded by NIH 2R01HL089897-06A1 and NIH 2R01HL089856-06A1. The authors would like to thank to all the COPDGene Investigators for their contributions.

5. REFERENCES

1. Bellamy, D., et al., *International Primary Care Respiratory Group (IPCRG) Guidelines: management of chronic obstructive pulmonary disease (COPD).* Prim Care Respir J, 2006. **15**(1): p. 48-57.
2. O'Reilly, J., et al., *Management of stable chronic obstructive pulmonary disease in primary and secondary care: summary of updated NICE guidance.* BMJ, 2010. **340**: p. c3134.
3. Feragen, A., et al., *Geometric tree kernels: classification of COPD from airway tree geometry.* Inf Process Med Imaging, 2013. **23**: p. 171-83.
4. Amaral, J.L., et al., *Machine learning algorithms and forced oscillation measurements to categorise the airway obstruction severity in chronic obstructive pulmonary disease.* Comput Methods Programs Biomed, 2015. **118**(2): p. 186-97.
5. Joshi, S. and H. Joshi. *SVM Based Clinical Decision Support System For Accurate Diagnosis Of Chronic Obstructive Pulmonary Disease*. in *International Journal of Engineering Research and Technology*. 2013. ESRSA Publications.
6. Glorot, X., A. Bordes, and Y. Bengio. *Domain adaptation for large-scale sentiment classification: A deep learning approach*. in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011.
7. Mohamed, A.-r., et al. *Deep belief networks using discriminative features for phone recognition*. in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. 2011. IEEE.
8. Raina, R., A. Madhavan, and A.Y. Ng. *Large-scale deep unsupervised learning using graphics processors*. in *Proceedings of the 26th annual international conference on machine learning*. 2009. ACM.
9. Regan, E.A., et al., *Genetic epidemiology of COPD (COPDGene) study design.* COPD: Journal of Chronic Obstructive Pulmonary Disease, 2011. **7**(1): p. 32-43.
10. Han, M.K., et al., *GOLD 2011 disease severity classification in COPDGene: a prospective cohort study.* The Lancet Respiratory Medicine, 2013. **1**(1): p. 43-50.
11. Duda, R.O., P.E. Hart, and D.G. Stork, *Pattern classification*. 2012: John Wiley & Sons.
12. Hinton, G.E., S. Osindero, and Y.W. Teh, *A fast learning algorithm for deep belief nets.* Neural Comput, 2006. **18**(7): p. 1527-54.
13. Mohamed, A.-r., D. Yu, and L. Deng. *Investigation of full-sequence training of deep belief networks for speech recognition*. in *INTERSPEECH*. 2010.
14. Bengio, Y., et al., *Greedy layer-wise training of deep networks.* Advances in neural information processing systems, 2007. **19**: p. 153.
15. Cai, D., C. Zhang, and X. He. *Unsupervised feature selection for multi-cluster data*. in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2010. ACM.
16. Lange, P., et al., *Prediction of the clinical course of chronic obstructive pulmonary disease, using the new GOLD classification: a study of the general population.* Am J Respir Crit Care Med, 2012. **186**(10): p. 975-81.
17. Burholt, V. and P. Nash, *Short Form 36 (SF-36) Health Survey Questionnaire: normative data for Wales.* J Public Health (Oxf), 2011. **33**(4): p. 587-603.
18. Hurst, J.R., et al., *Susceptibility to exacerbation in chronic obstructive pulmonary disease.* N Engl J Med, 2010. **363**(12): p. 1128-38.
19. Balcells, E., et al., *Factors affecting the relationship between psychological status and quality of life in COPD patients.* Health Qual Life Outcomes, 2010. **8**(108): p. 108.