Automated analysis of chest radiographs for cystic fibrosis scoring

Zhaowei Huang [1], Chen Ding[2], Lei Zhang[2], Min-Zhao Lee[1], Yang Song[1], Hiran Selvadurai[3], Dagan Feng[1], Yanning Zhang[2], Weidong Cai[1]

[1] Biomedical and Multimedia Information Technology (BMIT) Research Group, School of IT, University of Sydney, Australia
[2] Shaanxi Key Lab of Speech & Image Information Processing (SAIIP), School of Computer Science, Northwestern Polytechnical University, China
[3] Children's Hospital at Westmead, Sydney Children's Hospitals Network, Sydney, Australia
zhua7630@uni.sydney.edu.au

Abstract. We present a framework to analyze chest radiographs for cystic fibrosis using machine learning methods. We compare the representational power of deep learning features with traditional texture features. Specifically, we respectively employ VGG-16 based deep learning features, Tamura and Gabor filter based textural features to represent the cystic fibrosis images. We demonstrate that VGG-16 features perform best, with a maximum agreement of 82%. In addition, due to limited dimensionality, Tamura features for unsegmented images achieve no more than 50% agreement; however, after segmentation, the accuracy of Tamura can reach 78%. In combination with using the deep learning features, we also compare back propagation neural network and sparse coding classifiers to the typical SVM classifier with polynomial kernel function. The result shows that neural network and sparse coding classifiers outperform SVM in most cases. Only with insufficient training samples does SVM demonstrate higher accuracy.

Keywords: Cystic Fibrosis, computer-assisted score, deep learning feature, VGG-16

1 Introduction

Cystic fibrosis (CF) is a widespread life-threatening genetic disease, which affects up to 1 in 3000 people born in the highest-risk regions [1]. For example, Cystic Fibrosis Community Care* shows that 1 in 25 people in Australia are carrying defective CF genes and nearly 90 babies each year are born with this disease. The disease causes considerable morbidity and mortality, affecting multiple organs and ultimately with an average life expectancy at birth of close to 38 years despite ongoing medical care [2].

Cystic fibrosis causes major disease in the lungs. People with cystic fibrosis generally suffer from difficulty breathing, and frequent episodes of pneumonia. Half of patients with CF will require lung transplants. Clinicians usually assess the severity degree of cystic fibrosis by analyzing radiological images of the diseased lungs. For example, plain chest radiographs (CXRs) are often used to assess cystic fibrosis in

*www.cysticfibrosis.org.au/nsw/collaborative-research-project

children [3]. Shwachman-Kulczycki scoring is usually used in Australia to quantify the degree of abnormality in the lungs [4], with reference to the visible changes associated with the disease as seen on CXRs. In particular, clinicians look for signs of airflow obstruction (expanded shape of the chest cavity), bronchial and vascular thickening (linear markings), nodules and cysts, and gross regional abnormalities in lungs to give the assessment result. In this work, we mainly focus on the Shwachman-Kulczycki scoring system.

Shwachman-Kulczycki scoring classifies CXRs into five categories, which are quantified into a range from 5 to 25 with interval 5, in the order of decreasing severity. Table I describes the CXR findings for each score, as initially proposed by Shwachman and Kulczycki. Clinicians assign a Shwachman-Kulczycki score based on their own observations, which is a subjective determination and thus varies between different clinicians. Therefore, the development of an automatic scoring system providing clinicians with an objective measure of the CXR changes is still a challenging problem.

CF is an interstitial lung disease. The visual appearances of CF are mainly in the regional textures in the lungs. A recent study used Tamura, Gabor filter and other textural features to build a fully automated scoring of chest radiographs in cystic fibrosis and obtained 75% and 51% agreement with clinicians [5]. To the best of our knowledge, this is the best computer-assisted score for CF chest radiographs.

Table 1. Shwachman-Kulczycki X-Ray SCORING [4]

Grading	**Points**	**Findings**
Excellent	25	Clear lung fields.
Good	20	Minimal accentuation of bronchovascular markings; early emphysema.
Mild	15	Mild emphysema with patchy atelectasis; increased bronchovascular markings.
Moderate	10	Moderate emphysema; widespread areas of atelectasis with superimposed areas of infection; minimal bronchial ectasia.
Severe	5	Extensive changes in pulmonary obstructive phenomena and infection; lobar atelectasis and bronchiectasis.

Recently, neural networks, represented by Convolutional Neural Network (CNN), have shown excellent learning and classifying abilities. Moreover, some studies applied CNN to medical image processing [6,7,8,9]. The deep structure of neural networks enables the extraction of much more complicated features than the traditional textural features.

The purpose of our study is to build a framework for automated scoring with various feature extraction techniques and find more appropriate feature extraction methods and suitable classifiers to improve the accuracy and stability of the system. In contrast to previous methods considering textural features with support vector machine (SVM), this study proposes to employ deep learning methods to build an experimental system with deep learning features and deep learning classifiers for CXRs.

2 Methods

In order to ensure the correctness of the result, the proposed scoring framework consists of three steps, including the preprocessing, feature extraction and classification steps. In this paper, we use pre-existing fined-tune VGG-16 and seven-fold cross-validation to build up whole system (Fig 1).

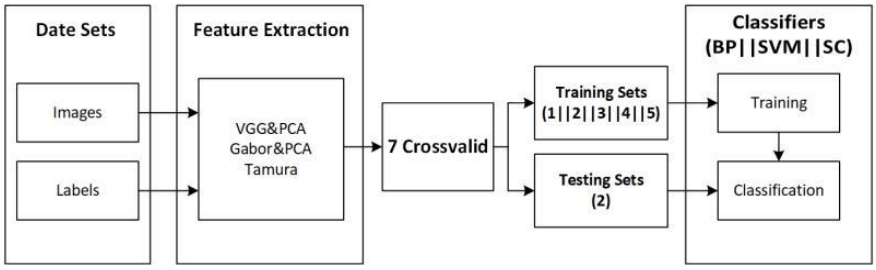

Fig. 1. The proposed scoring framework for cystic fibrosis in lungs

2.1 A. Acquisition of CXR Data

In this study, all experiment data come from the CXR data of 139 patients with cystic fibrosis, which are identified from an Australian pediatric cystic fibrosis registry and are aged between 2 to 16.

To evaluate the performance of scoring framework quantitatively, we consider the clinicians' reviewed results for all 139 images to be the standard score results. Out of all 139 images, clinical scoring assigned 36 images a score of 10, 56 images with a score of 15 and 47 images with a score of 20.

2.2 Preprocessing

CXR images were taken with different protocols and stored in several different formats. For each image, the preprocessing step includes edge clipping, resampling, and gray scale normalization.

In order to eliminate regions of the image outside the body, and to simplify resampling, we used the difference between the lung field and the background to get the axis-x and axis-y projection and crop out the external regions ("edge clipping"), as shown in Fig 2. The green lines are the local maximum for left or right, and the red line is the midline.

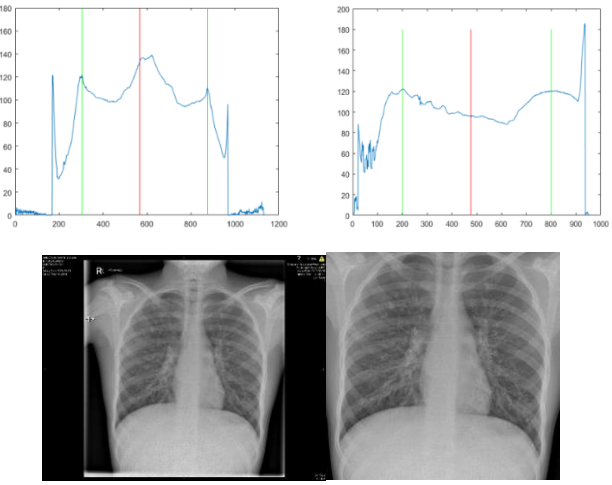

Fig. 2. Edge clipping: Horizontal projection, Vertical projection, original image, result,

Original CXR image dimensions ranged from 721 × 696 to 1131 × 951 pixels, with gray values between 0–30000. After edge clipping, images still contain the original depth for later experiments. To control the standard input dimension, we resample images to a size of 512×512 pixels, and gray levels scaled to 0–255. To retain discriminating information, we do not apply noise reduction or enhancement. We also perform automated segmentation [5], which we evaluated using overlap as our performance measure. The result was 0.939 in [5] and we achieve 0.899. The sample of segmentation results are listed in Fig 3.

Fig. 3. Segmented results

2.3 Feature Extraction

In the feature extraction step, we investigate the comparison between deep learning features and textural features. Since in the previous study, Tamura achieved the best performance and Gabor achieved fair results [5], we decided to use Tamura and Gabor as the textural features for comparison.

The Tamura features are based on psychophysical studies of the characterizing elements that are perceived in textures by humans: Contrast, Directionality, Coarseness, Linelikeness, Regularity, and Roughness. Among these, the first three are of greater importance. Contrast measures the way in which gray levels vary in the image and to what extent their distribution is biased to black or white. Directionality considers the edge strength and the directional angle. They are computed using pixel-wise derivatives according to Prewitt's edge detector. Coarseness relates to the distances of important spatial variations of grey levels, that is, implicitly, to the size of the primitive elements forming the texture. [10,11,12]

The Gabor filter is a linear filter that can be used for edge detection. It provides orientation selection and is biologically plausible [13]. The Gabor filter is generated by scaling and rotation from a parent wavelet so that it could extract the relevant features in different scales and directions in the frequency domain. We set 6 directions and 2 scales over each image, for a total of 12 filters to be used for feature extraction.

In contrast to the hand-crafted texture features mentioned above, deep learning features can be task-driven and learned from extensive training examples. More importantly, deep learning enables the learned feature to capture more flexible and complicated structure in the data, and thus improves the task performance. Witnessing the success of deep learning feature in various computer vision tasks, we turn to leverage deep learning features to represent abnormalities in cystic fibrosis.

Due to the limited data set, we adopt pre-trained deep learning features. Without generality, we choose the VGG-16 based deep learning features. The VGG-16 based deep learning feature set is pre-trained from the ImageNet database with 16 layers of deep convolutional neural network [14,15]. VGG-16 consists of 13 convolution layers and three fully connected layers. The VGG-16 based deep learning feature vector is the output of the last fully connected layer.

The VGG-16 network is pre-trained using an extensive collection of nature images. However, the cystic fibrosis CXR image is distributed completely differently from the nature images, which makes it difficult to depict cystic fibrosis well with the pre-trained VGG-16 feature. To address this problem, we employ the labeled cystic fibrosis images to fine-tune the pre-trained VGG-16 feature. Specifically, we modify the last fully connected layer to the size of the classification problem in cystic fibrosis. Then, we adjust the network parameters according to the classification error with backpropagation techniques. By doing this, the VGG-16 network can be adapted to fit the distribution of the cystic fibrosis images.

In this paper, we use the output of the 16th layer of the VGG-16 as a feature. This output contains 4096 nodes, so the feature has 4096 dimensions, and Gabor filter's output has 360 dimensions. To improve the computational efficiency and reduce the overfitting problem, we use the classical PCA dimensionality reduction algorithm to reduce the dimensions: the features from VGG-16 are reduced to 140-dimensions, and features from Gabor Filter were reduced to 60-dimensions.

2.4 Classifiers

We choose back propagation based neural network (BP) and Sparse Coding (SC) as classifiers and compare their performance with polynomial-kernel support vector machine (SVM).

BP is a neural network that uses error back propagation for training [16]. BP model is built up by three main element layers: input layer, hidden layer, and output layer. In hidden layer, the tunable parameters are the learning rate, number of iterations, and number of nodes number. In our method, we design the parameters as follows: learning rate: 0.0021; number of nodes in hidden layer: 850, 105 and 50; number of iterations: 10 (due to the initial parameters of the back propagation neural network.

SC [17] is an effective way of exploiting the data structure. In particular, through representing the data onto a given dictionary, SC exploits the underlying correlation among different data samples by depicting the sparsity on the representation. We set the lambda=0.01 using the cross-validation test.

SVM is a supervised learning method that has been widely used. It uses kernel functions to avoid the increase in computational complexity caused by increased dimension [18]. Due to the limited number of samples, we choose the polynomial kernel function. The values of the training parameters C and gamma are 1.8, determined by grid search and cross-validation test.

2.5 Cross-validation

In this step, we adopt cross-validation for validating the proposed method. Since the samples number is 139, we randomly choose one sample to copy and then add it to the dataset. The total 140 samples in the dataset can be evenly divided into seven groups. We randomly choose some groups for training and the other group for testing.

3 EXPERIMENT AND RESULTS

We conducted two experiments to check the classification effect of the deep learning features and deep learning classifiers, deep learning features（VGG-16） compared with textural features（Tamura, Gabor）; and deep learning classifiers (BP and SC) compared with machine learning algorithm（SVM), and the difference between using different numbers of training sets.

3.1 Deep learning classification performance verification

From randomly assigning 140 samples into seven groups of 20, we randomly selected two groups as a test set, and four groups from the remaining five groups as a training set for six different cycles. The results can be seen in Table 2.

Table 2. Comparison of classification performance

Feature	Classifier		
	BP	SC	SVM
VGG-16	0.82±0.06	0.80±0.05	0.79±0.08
Gabor	0.76±0.04	0.77±0.03	0.77±0.05
Tamura	0.48±0.02	0.41±0.04	0.42±0.05

From the perspective of features, whether it was in the deep learning classifiers or the SVM classifier，VGG-16's classification results were significantly better than the textural features. In back propagation neural network, VGG-16 achieved a result 6% higher than Gabor and 34% higher than Tamura. In sparse coding, VGG-16 was 3% better than Gabor and 39% better than Tamura. in SVM, VGG-16 was 2% better than Gabor and 37% better than Tamura. We also noticed that the variance of the VGG-16 feature is greater than the textural features. It is possible that for deep convolution neural networks such as VGG-16, the number of training samples in this paper is too small and pre-training is limited by the effect of parameter debugging.

From the perspective of classifiers, for both the deep learning features and traditional texture features that were used, the results of all three classifiers were similar. Overall, back propagation neural network classifier was slightly better than the other two classifiers. In VGG-16, BP classifier performed 2.5% better than both sparse coding and SVM; In Tamura, BP classifier was 7% better than sparse coding and 6% better than SVM; in Gabor, BP classifier was 1.3% below than other two classifiers. We also noticed that the variance of the deep learning classifier was less than that of the SVM classifier.

The results of this paper have a significant difference with [5]. First of all, the classification performance of Tamura features is obviously smaller than that of [5], which was 0.75. Second, even if the SVM classifier is used, the Gabor feature classification (0.77) is better than that of [5] (0.51). In this paper, the parameters of all three classifiers were optimized for the feature set after combining the three features, making the SVM core and parameter configuration was more reasonable. We also added an experiment to test the segmented images, with results shown in Table 3. It is obvious that segmentation is an important preprocessing step for Tamura.

Table 3. Comparison for Tamura

Feature	Classifier		
	BP	SC	SVM
Segmented	0.78±0.03	0.79±0.04	0.75±0.02
Unsegmented	0.48±0.02	0.41±0.04	0.42±0.05

VGG-16 and neural network produce 17.1% more agreement than an independent clinician observer (0.70). Even when combining VGG-16 and SVM, the results remain 12.9% better. Comparing with [5] (Tamura and SVM, 0.75), VGG-16 and neural

network produced 9.3% improvement; with VGG-16 and SVM, the results were 5.3% better.

3.2 Training set size optimization

From randomly assigning 140 samples into seven groups of 20, we randomly selected two groups as a test set, and between 1 and 5 groups from the remaining five groups as training sets to determine the effect of different training samples on the classifications.

Table 4 shows the relationship between classification performance, and training set size for VGG-16 features with different classifiers. Table 5 shows the relationship between classification performance and training set size for Gabor feature with different classifiers.

Table 4. Training sets size optimization for VGG-16

Training sets	Feature -Classifier		
	VGG-BP	VGG-SC	VGG-SVM
20	0.51±0.09	0.51±0.09	0.55±0.05
40	0.63±0.10	0.63±0.07	0.64±0.07
60	0.73±0.11	0.71±0.06	0.71±0.08
80	0.82±0.06	0.80±0.05	0.79±0.08
100	0.48±0.12	0.42±0	0.45±0

Table 5. Training sets size optimization for Gabor

Training sets	Feature -Classifier		
	Gabor -BP	Gabor -SC	Gabor -SVM
20	0.51±0.09	0.56±0.12	0.56±0.08
40	0.59±0.11	0.62±0.10	0.60±0.07
60	0.68±0.07	0.69±0.08	0.71±0.03
80	0.76±0.04	0.77±0.03	0.77±0.05
100	0.42±0.04	0.26±0	0.35±0

The result shows that, for both VGG-16 and Gabor features, the classification accuracy of the three classifiers increased with increasing training set size at first, reaching a maximum at a training set size of 80; but decreasing significantly when training set size increased to 100. One possible reason is that the number of experimental samples is small (140), and the characteristic dimension (140) and Gabor feature dimension (60) of VGG-16 are higher, resulting in overfitting.

Table 6 shows the results for Segmented Tamura features. It appears that Tamura achieves a better result than Gabor but is still lower than VGG-16.

Table 6. Training sets size optimization for Segmented Tamura

Training sets	Feature -Classifier		
	Tamura-BP	Tamura-SC	Tamura-SVM
20	0.51±0.09	0.49±0.10	0.50±0.08
40	0.62±0.12	0.65±0.09	0.67±0.05
60	0.77±0.07	0.71±0.03	0.71±0.02
80	0.78±0.03	0.79±0.04	0.79±0.02
100	0.50±0.12	0.47±0	0.40±0

4 Conclusion

In this study, we present an automated scoring system for chest radiographs (CXRs) in cystic fibrosis. In order to improve the performance of the computer-aided scoring system, we compare the effectiveness of various features and classifiers. The VGG-16 based neural network is fine-tuned to transfer the knowledge learned from extensive nature images classification to CF severity scoring, and ultimately results in an improved VGG network (modified-VGG) suitable for CF chest radiography. A three-layer back propagation neural network and sparse coding were used for classification. Through 7-fold cross validation training and test sample ratio optimization, a satisfactory score was obtained, and the best classification accuracy rate was up to 0.82. We have demonstrated that the CF chest radiograph scoring based on deep convolution neural network can obtain better accuracy than with normal textural features, with better agreement than independent clinician observer in some cases.

In the future, further experiments can be conducted on the following three aspects. First, the number of samples used in this paper is limited. We can increase the number of experimental samples to carry out more detailed and in-depth study to improve the accuracy and stability of the score. Second, we could go deeper by using ResNet-152, or investigate with mixture deep learning classifiers, and transfer learning methods [19, 20, 21]. Last, we can further investigate new finds in other medical applications [22, 23, 24].

References

1. Ratjen, F., Döring, G.: Cystic Fibrosis. In: Lancet, vol. 361, pp. 681-9 (2003)
2. Yankaskas, J. R., Marshall, B. C., Sufian, B., Simon, R. H., Rodman, D.: Cystic fibrosis adult care: consensus conference report. In: Chest, vol. 125 1 Suppl, pp. 1S-39S (2004)
3. Cleveland, R. H., Zurakowski, D., Slattery, D. M., Colin, A. A.: Chest radiographs for outcome assessment in cystic fibrosis. In: Proc. Am. Thorac. Soc., vol. 4, pp. 302-5 (2007)
4. Shwachman, H., Kulczycki, L. L.: Long-term study of one hundred five patients with cystic fibrosis. In: AMA J. Dis. Child., vol. 96, pp. 6-15 (1958)
5. Lee, M. Z., Cai, W., Song, Y., Selvadurai, H., Feng, D. D.: Fully automated scoring of chest radiographs in cystic fibrosis. In: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, pp. 3965-3968 (2013)

6. Li, Q., Cai, W., Wang, X., Zhou, Y., Feng, D. D., Chen, M.: Medical image classification with convolutional neural network. In: 13th International Conference on Control Automation Robotics & Vision (ICARCV), Singapore, pp. 844-848 (2014)
7. Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., et al.: Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? In: IEEE Transactions on medical image, VOL.35, NO.5, pp. 1299-1312 (2016)
8. Song, Y., Li, Q., Huang, H., Feng, D., Chen, M., Cai, W.: Low Dimensional Representation of Fisher Vectors for Microscopy Image Classification. In: IEEE Transactions on Medical Imaging, 36(8), pp. 1636-1649 (2017)
9. Orlando, J. I., Prokofyeva, E., Fresno, M. D., et al.: Convolutional neural network transfer for automated glaucoma identification. DOI: 10.1117/12.2255740 (2017)
10. Tamura, H., Mori, S., Yamawaki, T.: Textural Features Corresponding to Visual Perception. In: IEEE Trans. on Systems, Man, and Cybernetics, SMC-8, pp. 460-472 (1978)
11. Niblack, C. W., et al.: The QBIC project: querying images by content using color, texture, and shape. In: Proc. of SPIE, Storage and Retrieval for Image and Video Databases, Vol. 1908, San Jose, pp. 173-187 (1993)
12. Castelli, V., Bergman, L. D.: Image Databases: Search and Retrieval of Digital Imagery. In: Wiley: New York (2002)
13. Fogel, I., Sagi, D.: Gabor filters as texture discriminator. In: Biol. Cybernetics, vol. 61, pp. 103–113 (1989)
14. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., Li, F.: ImageNet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, pp. 248-255 (2009)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proc. International Conference on Learning Representations http://arxiv.org/abs/1409.1556 (2014)
16. Ding, C., Xia, Y., Li, Y.: Supervised segmentation of vasculature in retinal images using neural networks. In: International Conference on Orange Technologies, Xian, pp. 49-52. doi: 10.1109/ICOT.2014.6954694 (2014)
17. Schölkopf, B., Platt, J., Hofmann, T.: Sparse Representation for Signal Classification. In: Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference, 1, MIT Press, pp.609-616 (2007)
18. Cortes, C., Vapnik, V.: Support-vector networks. In: Machine Learning, vol. 20, pp. 273-297 (1995)
19. Noor, S. S. M., et al.: Hyperspectral Image Enhancement and Mixture Deep-Learning Classification of Corneal Epithelium Injuries. In: Sensors 17 (11), 2644 (2017)
20. Ren, J.: ANN vs. SVM: Which one performs better in classification of MCCs in mammogram imaging. In: Knowledge-Based Systems 26, 144-153 (2012)
21. Wang, X., et al.: ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: CVPR, pp. 3462-3471 (2017)
22. Zabalza, J., et al.: Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging. In: Neurocomputing 185, pp. 1-10 (2016)
23. Wang, Z., et al.: A deep-learning based feature hybrid framework for spatiotemporal saliency detection inside videos. In: Neurocomputing 287, pp. 68-83 (2018)
24. Noor, S. S. M., et al.: The properties of the cornea based on hyperspectral imaging: Optical biomedical engineering perspective. In: Systems, Signals and Image Processing (IWSSIP) (2016)