Interstitial lung disease detection in CT using an ensemble method of patch CNN and radiomic classifier

Adriel Silva de Araújo, Leonardo Amado, Dimitri B. Mantovani, Ana M. Marques da Silva, Marcio Pinho

Adriel Silva de Araújo, Leonardo R. Amado, Dimitri B. A. Mantovani, Ana M. Marques da Silva, Marcio S. Pinho, "Interstitial lung disease detection in CT using an ensemble method of patch CNN and radiomic classifier," Proc. SPIE 12464, Medical Imaging 2023: Image Processing, 1246438 (3 April 2023); doi: 10.1117/12.2654452

SPIE.

Event: SPIE Medical Imaging, 2023, San Diego, California, United States

Interstitial Lung Disease Detection in CT using an Ensemble Method of Patch CNN and Radiomic Classifier

Adriel Silva de Araújo*[a,b], Leonardo R. Amado[a], Dimitri B. A. Mantovani[b], Ana Maria Marques da Silva [b,c], Marcio S. Pinho[a]

[a]Graduate Program in Computer Science, School of Technology, PUCRS, Porto Alegre, RS, Brazil; [b]Medical Image Computing Laboratory (MEDICOM), PUCRS, Porto Alegre, RS, Brazil, [c]Medical Image & Data Analytics (MEDIIMA), San Diego, CA, USA

ABSTRACT

Interstitial Lung Disease (ILD) refers to pulmonary disorders that affect the lung parenchyma through inflammation and fibrosis. It is possible to diagnose ILD visually with computed tomography (CT), but it is highly demanding. Machine learning (ML) has yielded powerful models, such as convolutional neural networks (CNN), that achieve state-of-the-art performance in image classification. However, even with advances in CNN explainability, an expert is often required to justify its decisions adequately. Radiomic features are more readable for medical analysis because they can be related to image characteristics and are intuitively used by radiologists. There is potential in using image data via CNN and radiomic features to classify lung CT images. In this work, we develop two ML models: a CNN for classifying ILD using CT scans; and a Multi-Layer Perceptron (MLP) for classifying healthy and ILD using radiomic features. In the ensemble approach, output weights of each model are combined, providing a robust method capable of classifying ILD with the CT and the radiomic features. From a high-resolution CT dataset with 32 x 32 patches of pathological lung and healthy tissues, we extract 92 radiomic features, excluding those above 90% Pearson correlation in the training sets of both cross-validation and final models. Using 0.6 for the MLP and 0.4 for the CNN as weights, our approach achieves an accuracy of 0.874, while the MLP achieved 0.870 and, the CNN.

Keywords: Radiomics, Ensemble Classification, Computed Tomography, Machine Learning, Deep Learning.

1. INTRODUCTION

Interstitial Lung Disease (ILD) refers to pulmonary disorders that affect the pulmonary parenchyma through inflammation and fibrosis. ILD has many causes, including occupational and environmental exposure to toxins and ionizing radiation [1], [2]. It causes scar tissue, enabling the lungs to transport oxygen into the bloodstream effectively. A conservative incidence shows a range of 3–9 cases per 100,000 people per year in Europe and North America, with a lower incidence in East Asia and South America [3]. Early diagnosis is crucial for making treatment decisions, whereas a misdiagnosis may lead to life-threatening [4]. It is possible to diagnose ILD through a patient's family history, blood analysis, biopsies, and radiologic procedures like radiography and computed tomography (CT) [1,2].

CT imaging can accurately differentiate between types of ILD due to its capability to provide high spatial resolution volumetric images with multiple radiological patterns. These patterns represent structural alterations in pulmonary tissue and are defined by their shape and the distribution of the radiation attenuation. For example, emphysema, fibrosis, ground glass opacity, micronodules, and honeycombing are possible classifications that arise because of these characteristics, and their correct classification and spatial predominance in the lungs allow their association with specific diseases, which helps the differential diagnosis [5].

Artificial intelligence-based methods have not been significantly deployed in clinical practice, probably due to the underlying "black-box" nature of the deep learning algorithms increasing risks in diagnostics. Even with underlying statistical principles, there is a lack of ability to explicitly represent the knowledge for a medical image interpretation task performed by a deep neural network. However, classical machine learning (ML) techniques for image classification can learn via the information included in the radiomic image biomarkers, with each value directly correlated with a gray level intensity or texture characteristic of the images. Convolutional Neural Networks can classify patterns directly with image

Medical Imaging 2023: Image Processing, edited by Olivier Colliot, Ivana Išgum,
Proc. of SPIE Vol. 12464, 1246438 · © 2023 SPIE
1605-7422 · doi: 10.1117/12.2654452

data, and it is possible to explain their classifications with Activations Maps [6], for example, by identifying the most relevant image areas for the classification task. Such techniques can alleviate the inherent lack of explainability that affects many ML techniques, making them more accessible to medical teams.

Regardless of the explainability issues, CNNs are achieving state-of-the-art performance in classification using non-structured data and medical image classification [7]. Pattern identification on CT images plays a central role in diagnosing and treating patients with ILD. However, this task can be challenging even for trained radiologists. Thus, supporting tools, such as computer-aided diagnosis (CAD) systems for CT pattern detection and disease classification, can improve the diagnosis and lead to early treatment of pulmonary diseases. There is potential in using image data and radiomic features to classify lung images, and each technique has its pros and cons. Since each method uses different input data and the models have other inductive biases, we propose a novel method for the ILD classification using CT imaging. In this work, we develop two ML models: a CNN for classifying ILD using image data from CT scans; and a model using classical ML for classifying ILD using radiomic features. Finally, we combine both in an ensemble capable of leveraging the strengths of both methods. The ensemble method weights the output of each model, providing a robust method capable of classifying ILD through CT images.

2. METHODS

Images were obtained from a collection built at the University Hospitals of Geneva (HUG) [5]. The dataset contains high-resolution CT series with annotated small regions or patches of pathological lung tissue along with clinical parameters, publicly available at (http://medgift.hevs.ch/wordpress/databases/ild-database/). A total of 18784 patches of 32x32 pixels, with 8 bits/pixel each, were separated per patient into training (80%) and testing (20%) between healthy and non-healthy classes. Figure 1 shows representative patches from non-healthy (A) and healthy (B) classes.

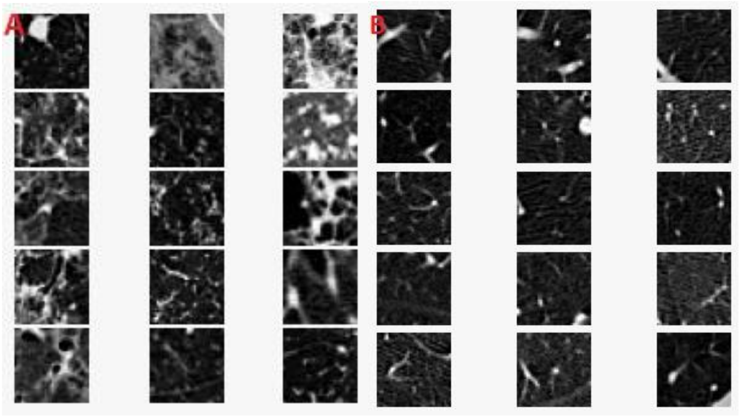

Figure 1. Representative patches from non-healthy (A) and healthy (B) classes.

For each patch, we extracted 92 radiomic features using PyRadiomics 2.2.0 [8], an open-source package for radiomics in Python 3.7, without any previous application of filtering. Ninety-two features were extracted, including first-order statistics, gray level co-occurrence matrix (GLCM), gray level run length matrix (GLRLM), gray level size zone matrix (GLSZM), neighboring gray-tone difference matrix (NGTDM), and gray level dependence matrix (GLDM).

2.1 Study Design and Architecture

We elaborated a methodology for the classification based on the combination of a CNN image classifier and an MLP radiomics classifier. Figure 2 shows the classification design scheme.

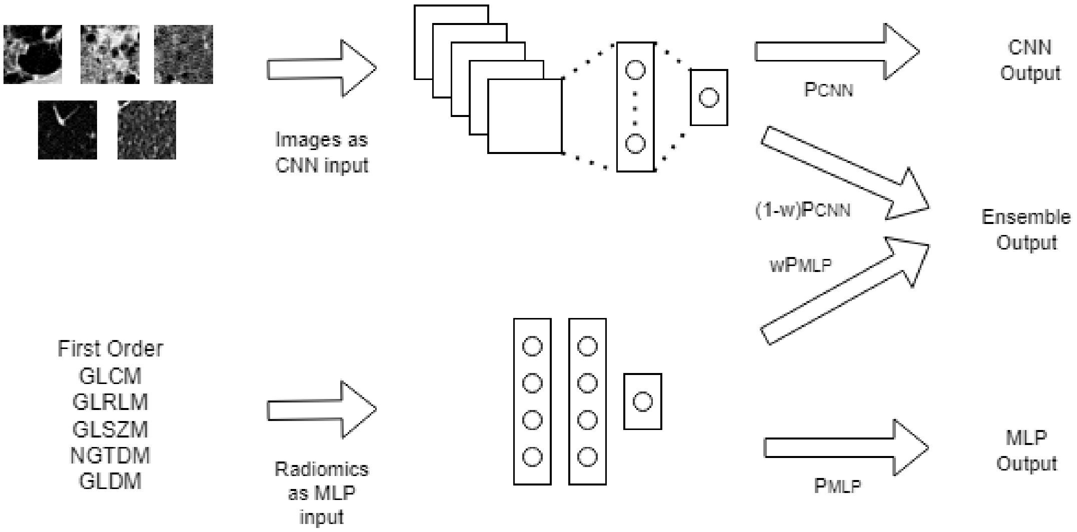

Figure 2. Design scheme for the ILD classification mechanism. Both an image and a radiomics classifier are used to perform the prediction.

We developed an MLP to classify radiomic features with two hidden layers, ten neurons each, with batch normalization and dropout (0.5) for regularization. Also, we developed a CNN for image classification. The model contained 6 convolutional layers, 80 filters each, with padding so that the convolutional volume does not shrink, and a final 50-neuron dense layers, with batch normalization and dropout (0.4) for regularization. Both models used Rectifier Linear Units for intermediate activations, the Sigmoid for the last activation, and binary cross-entropy as the loss function, Adaptative Momentum Estimation, using momentum terms 0.9 and 0.999 and initial learning rate of 0.001, as the gradient descent optimizer, and the training regime occurring in 50 epochs max, using early stopping with 15 epochs of tolerance. We performed all neural network manipulation using TensorFlow 2.8.2. Both MLP and CNN have a probabilistic output between 0 and 1. We modulated these outputs with weights so that $wP_{MLP} + (1-w)P_{CNN}$,â • 0.5, where w is a particular weight for the models, PMLP is the probabilistic output of the MLP, and PCNN is the probabilistic output of the CNN. We chose w via validation.

As for data pre-processing, images were divided by 255 and modified via random rotations, flips and contrast modifications using the Albumentations 1.2.1 library [9]. For the radiomics data, from the 92 calculated features, we excluded those with an above 90% pairwise absolute value for Pearson correlation in the training and the validation datasets.

2.2 Validation

We performed 10-fold cross-validation, separating the training data into 10 subsets, where we inspected the performance of such models. We ensured that data from the same subject were not simultaneously in the training and validation. We iterated the value for w between 0 and 1, with increments of 0.1, and selected the value that maximized the accuracy.

Table 1 shows the descriptive statistics for the cross-validation procedure. Each row shows the mean and standard deviations of the evaluation metrics during the 10 rounds for a particular value of w, sorted by the highest accuracy. It is noted that the value of w = 0.6 returns the best performance for the ensemble classifier. The performance of w = 0 (only CNN) and w = 1 (only MLP), are located in the last rows of Table 1. The mean values for the number of epochs before stopping the training were 45 and 25 for the MLP and CNN, respectively. These numbers were used to train the final models.

Table 1. Accuracy and F1-score values for each w-value in validation data.

Mean Accuracy	Accuracy St.dev	Mean F1	F1 St.dev	Weight (*w*)
0.816	**0.113**	**0.629**	**0.258**	**0.6**
0.812	0.116	0.628	0.267	0.7
0.809	0.120	0.619	0.242	0.5
0.806	0.116	0.619	0.271	0.8
0.803	0.123	0.611	0.237	0.4
0.800	0.118	0.612	0.273	0.9
0.800	0.125	0.604	0.236	0.3
0.796	0.124	0.592	0.234	0.2
0.794	0.120	0.604	0.274	1
0.790	0.125	0.581	0.225	0.1
0.785	0.125	0.567	0.222	0

3. RESULTS AND DISCUSSION

Using all data from the training set, we evaluated the performance of the models in the test set. Table 2 shows accuracy and F1-Score values for the test data using an MLP or CNN alone, as well as an ensemble of the models.

Table 2. Models' metrics for test data.

Model	Test accuracy	Test F1-Score
MLP	0.870	0.809
CNN	0.847	0.836
Ensemble	0.874	0.798

The higher performance of the ensemble model can be explained by the independence of the features learned by the CNN and the radiomic biomarkers used by the MLP. CNN filters learn non-linear, higher dimensional features of the convolutional volumes, which are pre-processed by the previous layers in the network. Radiomic biomarkers represent pixel statistics, handcrafted to inspect the intensity and texture characteristics of the images directly. As the CNNs become deep, the complexity of the features they learned increases, making the last dense layer a representation of those features learned by the various hidden convolutional layers. By describing visible, readily interpretable pixel statistics directly related to the radiologist's intuitive diagnostic technique, radiomic biomarkers introduce a higher-level analysis of the lower, non-linear CNN features, contributing to the classification.

The results show that the classifiers perform better when utilized together for the classification. However, the metrics' standard deviations are high compared to the average values. This can be explained by the fact that the pathological lung tissue patches are actually made of four subclasses: emphysema, fibrosis, ground-glass opacity, and micronodules [8].

Therefore, while correct in the pipeline, separating the dataset in a subject-by-subject manner may shift the unhealthy class's internal distributions to be biased towards a particular subclass or set of subclasses. This can lead to models performing better or worse depending on how much of the unhealthy patches are populated by a particular subclass.

Analyzing the increment of 3% in accuracy, when comparing the best and worse results in accuracy, we can see that, for a test size of 3757 patches, we can correctly classify 113 more patches of lung tissue. This number corresponds to 1 or even 2 subjects. However, it is vital to discuss that even though we can perceive the tendency of improvement with an ensemble classifier, the results' averages and standard deviations can improve by upgrading the model architectures. Especially, CNN architectures benefit from a more significant number of hyperparameters to obtain more non-linear features and, consequently, get better results. Smaller networks, like the ones analyzed in this work, may not have peak performance and have equal or worse evaluation metrics than a handcrafted feature classifier.

4. CONCLUSIONS

We evaluated the performance of an ensemble of a CNN and an MLP in the classification of lung CT patches and their radiomic features in healthy and unhealthy patterns. Some limitations and remarks about this work involve using feature selection by correlation, which is a more straightforward method. Still, it is model-independent, which may select some features that are not as relevant for that specific inductive bias. Another remark is that the science of artificial intelligence is a rapidly evolving field in which model architecture is constantly improving, with the contribution of attention models and vision transformers, for example. The idea here was to explore how different inductive biases could improve the classification of ILD CT images. As a future perspective, we will explore if these mixed classifiers have improved performance if we train a siamese, multi-input network instead of an ensemble of networks, as well as evaluate the performance of the ensemble methods in the multiclass ILD classification problem.

5. REFERENCES

[1] O. Kalchiem-Dekel, J. R. Galvin, A. P. Burke, S. P. Atamas, and N. W. Todd, “Interstitial Lung Disease and Pulmonary Fibrosis: A Practical Approach for General Medicine Physicians with Focus on the Medical History.,” *J. Clin. Med.*, vol. 7, no. 12, Nov. 2018, doi: 10.3390/jcm7120476.

[2] A. Wallis and K. Spinks, “The diagnosis and management of interstitial lung diseases,” *BMJ*, vol. 350, no. may07 17, pp. h2072–h2072, May 2015, doi: 10.1136/bmj.h2072.

[3] J. Hutchinson, A. Fogarty, R. Hubbard, and T. McKeever, “Global incidence and mortality of idiopathic pulmonary fibrosis: A systematic review,” *Eur. Respir. J.*, vol. 46, no. 3, pp. 795–806, Sep. 2015, doi: 10.1183/09031936.00185114.

[4] J. C. Muhrer, “Risk of misdiagnosis and delayed diagnosis with COVID-19: A Syndemic Approach,” *Nurse Pract.*, vol. 46, no. 2, p. 44, Feb. 2021, doi: 10.1097/01.NPR.0000731572.91985.98.

[5] A. Depeursinge, A. Vargas, A. Platon, A. Geissbuhler, P.-A. Poletti, and H. Müller, “Building a reference multimedia database for interstitial lung diseases.,” *Comput. Med. Imaging Graph.*, vol. 36, no. 3, pp. 227–38, Apr. 2012, doi: 10.1016/j.compmedimag.2011.07.003.

[6] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” *arXiv Prepr. arXiv1312.6034*, 2013.

[7] S. S. Yadav and S. M. Jadhav, “Deep convolutional neural network based medical image classification for disease diagnosis,” *J. Big Data*, vol. 6, no. 1, pp. 1–18, Dec. 2019, doi: 10.1186/S40537-019-0276-2/TABLES/16.

[8] J. J. M. van Griethuysen *et al.*, “Computational Radiomics System to Decode the Radiographic Phenotype,” *Cancer Res.*, vol. 77, no. 21, pp. e104–e107, Nov. 2017, doi: 10.1158/0008-5472.CAN-17-0339.

[9] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, “Albumentations: Fast and Flexible Image Augmentations,” *Inf. 2020, Vol. 11, Page 125*, vol. 11, no. 2, p. 125, Feb. 2020, doi: 10.3390/INFO11020125.