IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 3, 2019, accepted May 4, 2019, date of publication May 22, 2019, date of current version May 30, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2916849

A Novel Approach for Multi-Label Chest X-Ray Classification of Common Thorax Diseases

IMANE ALLAOUZI AND MOHAMED BEN AHMED
LIST/FSTT, Abdelmalek Essaadi University, Tangier 90040, Morocco

Corresponding author: Imane Allaouzi (imane.allaouzi@gmail.com)

ABSTRACT Chest X-ray (CXR) is one of the most common types of radiology examination for the diagnosis of thorax diseases. Computer-aided diagnosis (CAD) was developed to help radiologists to achieve diagnostic excellence in a short period of time and to enhance patient healthcare. In this paper, we seek to improve the performance of the CAD system in the task of thorax diseases diagnosis by providing a new method that combines the advantages of CNN models in image feature extraction with those of the problem transformation methods in the multi-label classification task. The experimental study is tested on two publicly available CXR datasets ChestX-ray14 (frontal view) and CheXpert (frontal and lateral views). The results show that our proposed method outperformed the current state of the art.

INDEX TERMS CAD, CXR, transfer learning, CNN, computer vision, multi-label classification, problem transformation method, deep learning, image classification, image feature extraction, thoracic pathologies.

I. INTRODUCTION

The thorax also called chest, is the upper part of the trunk located between the neck and the abdomen. It is mostly protected and supported by the rib cage, spine, and shoulder girdle. The rib cage is bounded by neighboring ribs and muscles and contains viscera, mainly the lungs, heart, and mediastinum organs, which have a vital role in feeding (esophagus), breathing, and pumping the blood to all parts of the body.

Chest pain is the most frequent reason for consultation and emergency room visits. Chest radiography, colloquially called Chest X-Ray (CXR), is one of the most common types of radiology examination for the diagnosis of thorax diseases. However, radiology involves decision-making under conditions of uncertainty, and therefore cannot always produce infallible interpretations or reports [1]. In this purpose, Computer-Aided Diagnosis (CAD) was developed to help radiologists to achieve diagnostic excellence in a short period of time and to enhance patient healthcare. CAD systems are not meant to replace or compare with doctors, but they are used as a ‘‘second opinion’’ complementary to that of a radiologist.

Over the past few years, a lot of interest and attention has been paid to improve CAD systems using Artificial Intelligence (AI) and Computer Vision (CV) techniques. One of the core problems and the typical task is medical image classification. The intent of the classification process is to assign single or multiple diagnostic outcomes to a medical image based on its content. In this regard, many efforts have been made to develop advanced classification approaches and methods in order to improve classification accuracy.

Originally, a dual-stage approach was used to tackle image classification problem. Where the first stage aims to extract hand-crafted features from image using feature descriptors, and then the extracted features are provided as input to a trainable classifier in the second stage [2]. However, the accuracy of this approach depends highly on the method used for feature extraction in the first stage. For this reason, deep learning was investigated in the task of image classification; it allows automatic extraction of features and classification by modeling data through multiple processing layers containing non-linearity.

The Convolutional Neural Networks (CNNs) are the most favorite and popular deep learning models for the task of image classification since it provides high accuracy and impressive results compared with other models. It was specially designed for use on two-dimensional data, such as image and video. The first CNN model was proposed in the late 90s, its basic idea is inspired from the human visual perception of recognizing things. Of these, the best known is the LeNet architecture that was used to read zip codes, digits, etc. [3].

However, these models are immensely data-hungry and rely on huge amounts of labeled data to achieve their performance, which is one of the most important obstacles since

The associate editor coordinating the review of this manuscript and approving it for publication was Shuihua Wang.

labeled data are not always available, in some tasks and domains, or are expensive to obtain. Therefore, a number of research studies have introduced the use of transfer learning method, where what has been learned in one setting is exploited to improve generalization in another [4].

Nowadays, multi-label classification methods are receiving increased interest and are required by many real-world application domains such as text classification and protein function classification. In multi-label classification, each instance can be assigned simultaneously into multiple classes. This is performed by either transforming the problem into one or more single-label sub-problems or by adapting a single-label classifier to handle multi-label data directly.

This paper presents our contribution to the task of detecting thoracic diseases from chest X-ray images using transfer learning and multi-label problem transformation methods. The main idea is to extract relevant features from CXRs using a pre-trained CNN and then classify the extracted features with multi-label problem transformation methods that transform the multi-label problem to single-label classification. The experimental study is tested on two publicly available datasets for CXRs. The results show that our proposed method outperformed previous works and introduce a new state-of-the-art with an average AUC of 0.882.

II. RELATED WORK

Various works have been proposed to automatically classify thorax diseases from frontal CXRs, thanks to the public release of the ChestX-ray14 dataset [19]: In the work of [19], they evaluated four classic CNN architectures to classify and localize disease lesion areas in a weakly supervised manner. In order to exploit label dependencies, [20] presented a two-stage end-to-end neural network model that combines a densely connected image encoder with a recurrent neural network decoder. While [21] investigate that which loss function is more suitable for training CNNs from scratch and present a boosted cascaded CNN for global image classification. The well-known of these works are CheXNet [22] that fine-tunes a DenseNet-121 on the global chest X-ray images, which has a modified last fully-connected layer and [23] that proposes an attention guided two-branch convolutional neural network for thorax disease classification. The proposed network is trained by considering both the global and local cues informed in the global and local branches respectively, and has achieved superior performance over the state-of-the-art approaches on CXR dataset.

Unlike the above works, a novel DualNet architecture [24] was introduced to assess frontal, as well as lateral CXRs. It emulates routine clinical practice by taking into account both view types simultaneously.

III. TRANSFER LEARNING WITH MEDICAL IMAGE

Transfer learning was used in the medical domain to face the problem of scarce and insufficient annotated images, and also to reduce the efforts of building a model from scratch for a specific task. Generally speaking, transfer learning aims to exploit knowledge present in a large training data from a source domain, and then use it either as initialization or a fixed feature extractor to enhance a model's performance in a target domain.

Transfer learning using CNN is a commonly used strategy to tackle the task of medical image classification, where a pre-trained CNN on a huge dataset (such as ImageNet [5]) can be exploited in three ways:

A. CNN ARCHITECTURE

In this case, we are just interested in the architecture. So, we have to train our model from scratch by fine-tuning all the layers of the CNN according to our dataset again. Instead of defining random weights as a starting point, we can use initial weights obtained from the pre-trained network that gives a good starting point against random initialization of the weights.

B. FREEZING CNN LAYERS

In this case, we freeze some of the earlier layers of the pre-trained CNN (due to overfitting concerns) which contain general features and only fine-tune some higher layers of the network that contain more specific features related to the properties of classes contained in the original dataset.

C. CNN AS FEATURES EXTRACTOR

In this case, we get rid of the last or latest fully-connected layers of a pre-trained CNN on a huge dataset in order to use the entire network as a fixed feature extractor. Then we train a linear classifier such as softmax classifier or SVM to predict label for a new dataset.

To tackle our problem of multi-label CXR classification, we chose to follow the first and the last strategy. First we train a pre-trained CNN on a huge dataset from scratch by fine-tuning all layers, then we remove the last fully connected layer (that predicts the diagnostic outcomes) and finally, we treat the rest of CNN only as a fixed feature extractor for our CXR dataset.

In order to choose the best CNN model for our task, we had to choose between several CNNs that have been applied for CXR classification including ResNet [6], VGG-Net [7], and DenseNet [8]. As a result, we chose the DenseNet-121 model which achieved the state-of-the-art results.

IV. MULTI-LABEL CLASSIFICATION (MLC)

Recently, Multi-label image classification has gained a surging interest in the field of computer vision and has been applied to tackle the problem of image and video annotation. Unlike single-label (binary/multi-class) image classification, where each image has only one label, a multi-label classifier can assign to an image multiple labels, exactly one or no label at all.

Different approaches have been proposed to address the problem of multi-label classification; they are mainly arranged into three categories:

Problem Transformation Methods: The main idea is to fit data to an algorithm by transforming the multi-label classification problem into one or more single-label (binary/multi-class) sub-problems, and then combine their results to form the multi-label prediction. Representative algorithms include Binary Relevance [9], Random k-Labelsets [10] and Classifier Chains [11].

Problem Adaptation Methods: The core idea is to fit an algorithm to data by extending popular learning techniques to deal directly with multi-label data. Representative algorithms include an adaptation of lazy learning techniques ML-kNN [12], an adaptation of decision tree techniques ML-C4.5 [13], an adaptation of kernel techniques Rank-SVM [14].

A novel approach called *Ensemble methods* [15] was developed on top of these two approaches. It consists in transforming the problem of multi-label classification into an ensemble of multi-label sub-problems. Representative algorithms include the Random k-labELsets method (RAkEL) [16], Ensemble Classifier chains (ECC) [17], and label space partitioning classifiers [18].

In order to find the most suitable approach for our case, we tried to make a comparison between the three above methods. However, both problem adaptation methods and ensemble methods have high memory requirements and take a considerable amount of time (>20 hours) when it is run on our dataset. For this, this work is carried out using problem transformation methods including Binary Relevance, Label Powerset, and Classifier Chains.

Binary Relevance (BR): This method assumes that labels are independent; it converts the multi-label task into k binary classification problems where k is the number of labels. So, it creates k datasets and train k binary classifier on each of these datasets. For a new instance, each of k binary classifiers votes separately to get the final result. The main advantages of BR are: low computational complexity and high flexibility as labels can be added or removed without affecting the model. However, BR suffers from two problems: the first one is that BR ignores interdependences between labels and the second one is the data imbalance that may occur after the transformation.

Label Powerset (LP): This method assumes labels are dependent; it converts the multi-label task into multi-class classification problem with k different classes where k is the number of possible combination of labels. This method is straightforward. Nevertheless, the main drawbacks of this method are: the computational cost is exponential with the original label set, and after the transformation, it is possible to have limited training examples for classes with less frequent combinations, producing data imbalance problem.

Classifier Chain (CC): As BR, CC transforms a multi-label problem into k binary classification problems where k denotes a set of labels and for each label; a separate binary classifier is designed. Classifiers are linked along a chain where the input for each classifier is different. This chaining method passes label information between classifiers, allowing CC to take into account label correlations and thus overcoming the label independence problem of BR. However, CC still retains the advantages of BR.

V. THE PROPOSED APPROACH

The idea behind our approach is to combine the effectiveness of CNN for image features extraction from a small image dataset and the power of the problem transformation methods in the task of multi-label classification. As shown in Figure 1, the development of the proposed method consists of four parts: data description and exploration, data pre-processing, feature extraction part, and classification part.

A. DATA DESCRIPTION AND EXPLORATION

1) CHESTX-RAY14 DATASET

The dataset contains 112,120 frontal CXRs from 30,805 unique patients. All CXRs are PNG format and have a size of 1024 × 1024. CXRs are labeled with 14 common thorax diseases including Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural_thickening, Cardiomegaly, Nodule, Mass and Hernia. If none of these diseases has been detected in a CXR, then it will be labeled as ''No finding''.

Visual exploration is an important step that allows us to understand what is in a dataset and the characteristics of the data including the size, format, and distribution of data. This is illustrated in the following figures.

2) CHEXPERT DATASET

CheXpert [25] is a large public dataset for chest radiograph interpretation, comprises 224,316 CXRs of 65,240 patients. Both frontal and lateral CXRs (see Figure 4) have been retrospectively collected from Stanford Hospital, performed between October 2002 and July 2017 in both inpatient and outpatient centers. However, in this study, we have worked only with 134,327 CXRs (115723 of frontal views and 18604 of lateral views) because we have ignored CXRs with uncertain labels. Each CXR is labeled with one or more pathology labels including Atelectasis, Cardiomegaly, Enlarged Cardiomediastinum, Consolidation, Pneumonia, Pneumothorax, Edema, Lung opacity, Lung Lesion, Pleural Effusion, Pleural other, Fracture, Support devices, No finding.

To better understand our dataset, we counted the number of CXRs for each pathology label and the number of CXRs with multiple labels (see Figure 5 and Figure 6).

B. DATA PRE-PROCESSING

Data pre-processing is meant to adequate our CXRs to the format the pre-trained model requires so we have resized CXRs to the required size 224 × 224 pixels. And in order

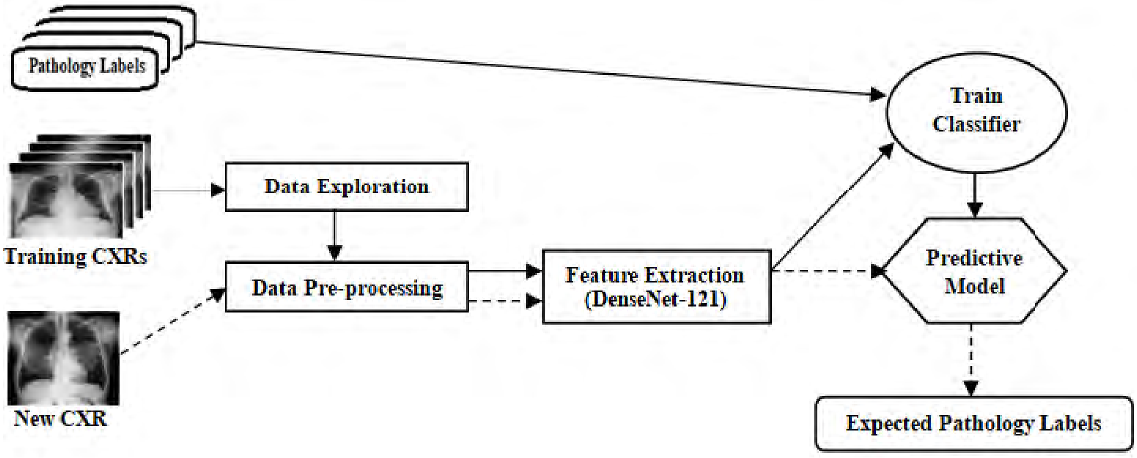

FIGURE 1. Multi-label CXR classification pipeline.

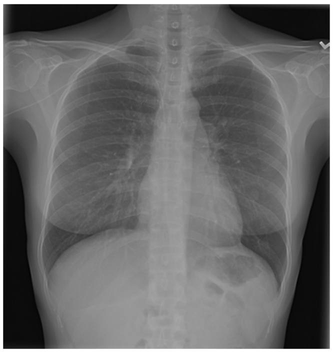

FIGURE 2. Example of ChestX-ray14 dataset: frontal CXR of a patient with infiltration and nodule.

to augment the dataset and make convergence faster while training the network, we utilized horizontal flipping and normalized our data by subtracting the mean from each pixel and then dividing the result by the standard deviation.

C. FEATURE EXTRACTION

The main goal of this phase is: given a CXR, generate the features that will subsequently be fed to a classifier in order to classify the CXR into one or multiple possible classes. To this end, a denseNet-121 model is used as a feature extractor.

The Dense Convolutional Network (DenseNet) [8] is a new CNN architecture that has outperformed the state-of-the-art results on most highly competitive object recognition benchmark tasks. The core idea of DenseNet is to ensure maximum information flow between layers in the network by connecting all layers (with matching feature-map sizes) directly with each other. As shown in Figure 8, this introduces $\frac{L\times(L+1)}{2}$ connections in an L-layer network, instead of just L, as in traditional architectures.

A DenseNet is a stack of dense blocks followed by transition layers. A dense block consists of a series of units. Each unit packs two convolutions, each preceded by Batch Normalization and ReLU activations. In addition, each unit generates a fixed number of feature vectors. This parameter, called growth rate, controls the amount of new information that layers can transmit. The layers between these dense blocks are transition layers which perform down-sampling of the features passing the network. A detailed explanation of DenseNet-121 architecture, the DenseNet we used in this work, is shown in Table 1.

TABLE 1. The DenseNet-121 architecture.

Layers	Output Size	DenseNet-121
Convolution	112x112	7x7 conv, stride2
Pooling	56x56	3x3 max pool, stride 2
Dense Block 1	56x56	[1x1 conv 3x3 conv] x 6
Transition Layer 1	56x56	1x1 conv
	28x28	2x2 average pool, stride 2
Dense Block 2	28x28	[1x1 conv 3x3 conv] x 12
Transition Layer 2	28x28	1x1 conv
	14x14	2x2 average pool, stride 2
Dense Block 3	14x14	[1x1 conv 3x3 conv] x 24
Transition Layer 3	14x14	1x1 conv
	7x7	2x2 average pool, stride 2
Dense Block 4	7x7	[1x1 conv 3x3 conv] x 16
Classification Layer	1x1	7x7 global average pool
	14D fully connected, sigmoid	

Motivated by the results obtained by DenseNet-121 on ChestX-ray14 dataset [23], [24], we have trained the DenseNet-121 model on our dataset, using initial weights

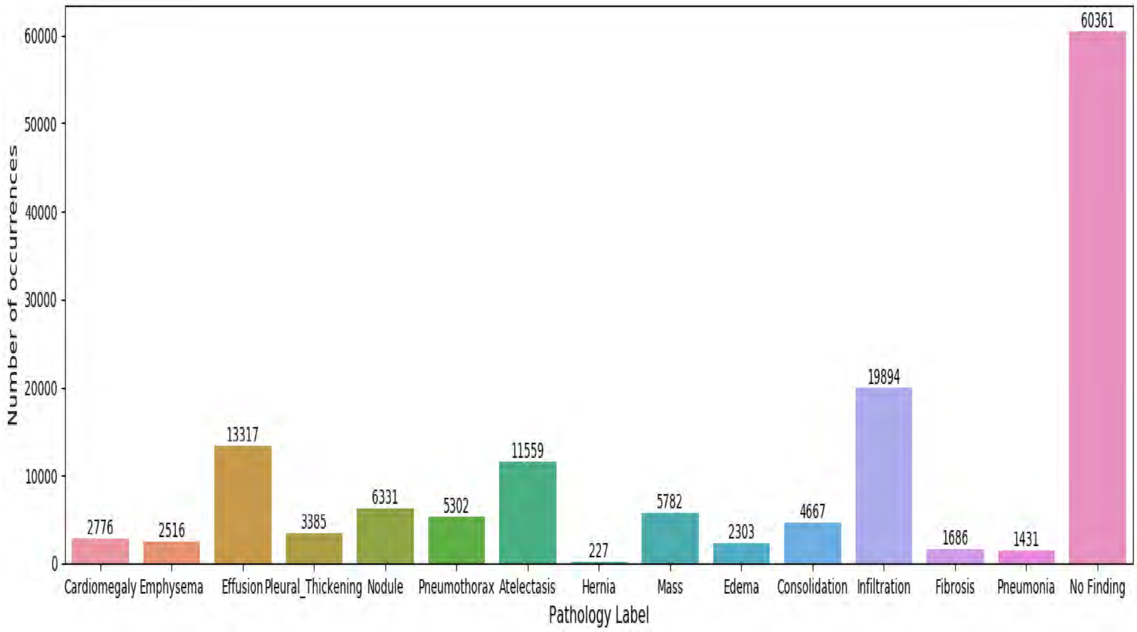

FIGURE 3. Number of CXRs, in ChestX-ray14 dataset, per pathology label.

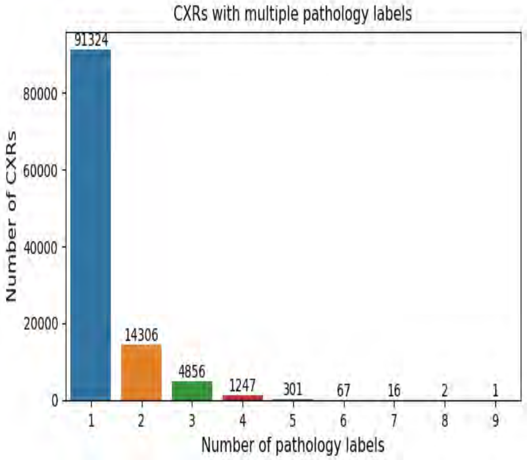

FIGURE 4. Number of CXRs, in ChestX-ray14 dataset, having multiple pathology labels.

obtained from the pre-trained network, on ImageNet, which gives a good starting point against random initialization of the weights. We have used a mini-batch size of 8 samples, and a number of epochs up to 110, the binary cross-entropy as a loss function where the best model was selected based on the validation loss. The Adam optimizer is used with an initial learning rate of 0.001 which is multiplied by 10 each time the validation loss plateau after an epoch.

Since this trained model is used only as feature extractor, we have removed the last fully connected layer (14D fully connected, sigmoid) and obtained a fixed feature vector of 1024D.

D. CLASSIFICATION

After the feature extraction stage, a classifier is needed to find the corresponding label(s) for every CXR. This is carried out using problem transformation methods including Binary Relevance, Label Powerset and Classifier Chains. Commonly used base classifier algorithms with these methods are SVM, J48 and Logistic Regression (LR). In this work, we choose to work with LR since it is the fastest and the more accurate for our specific task.

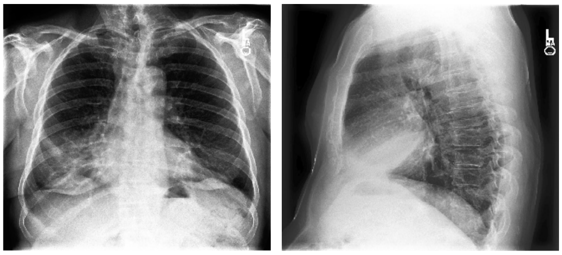

FIGURE 5. Example of CheXpert dataset: Frontal and lateral CXR of a patient with Pneumonia.

LR is a linear classifier. Its basic form seeks a hyperplane that separates data belonging to two classes. A brief description is as follows: assuming we have the set of training data (x1,y1),(x2,y2),…(xn,yn) and we want to classify the set into two classes where xi ϵ Rd is the feature vector and yi ϵ {0,1} is the label class. These classes are separated by a hyperplane wx + b = 0, where the conditional probability for LR classifier takes the following form:

$$p(y_i|\mathrm{x_i}) = \frac{\mathrm{e}^{((w\mathrm{x_i}+b)y_i)}}{1+\mathrm{e}^{((w\mathrm{x_i}+b)y_i)}}, \quad \mathrm{i}=1\ldots\mathrm{n} \tag{1}$$

A probability close to 1 means xi is very likely to be part of that label. The classifier parameters w and b can be determined by minimizing the average logistic loss function:

$$l_{avg}(w,b) = \frac{1}{m}\sum_{i=1}^{n}\log(1+e^{((wx_i+b)y_i)}) \tag{2}$$

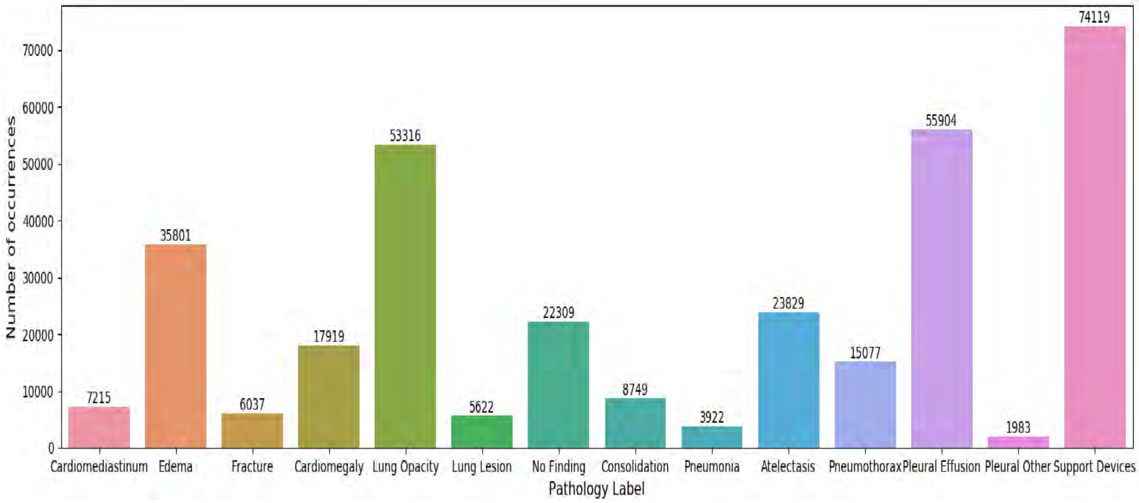

FIGURE 6. Number of CXR, in CheXpert dataset, per pathology label.

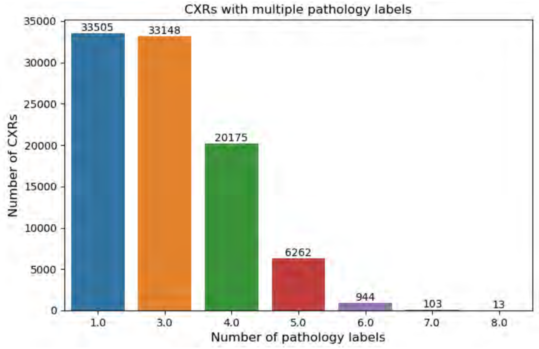

FIGURE 7. Number of CXRs, in CheXpert dataset, having multiple pathology labels.

$$\arg\max_{w,b} l_{avg}(w, \mathrm{b}) \tag{3}$$

VI. EVALUATION METRICS FOR MLC TASK

Unlike the traditional classification problems, where the prediction can be either correct or wrong, the multi-label classification problem is a more challenging task and requires more special evaluation measures since the performance over all labels should be taken into account. In a multi-label classification problem, a prediction can be fully correct (all predicted labels are correct), partially correct (some of the predicted labels are correct) or fully wrong (all predicted labels are wrong).

The evaluation metrics of MLC are broadly grouped into two categories:

A. EXAMPLE-BASED METRICS

The main idea is to first evaluate the average difference between the predicted and ground-truth classes for each test examples, and then average over all examples in the test set. The commonly used Example-Based metrics to evaluate the performance of a multi-label classification model are: Hamming Loss (HL), which reports how many times on average, an example-label pair is misclassified and the well-known metrics from the Information Retrieval (IR) Recall (R), Precision (P) and F1-score that combines the precision and recall measures of a classifier by means of an evenly harmonic mean of both them.

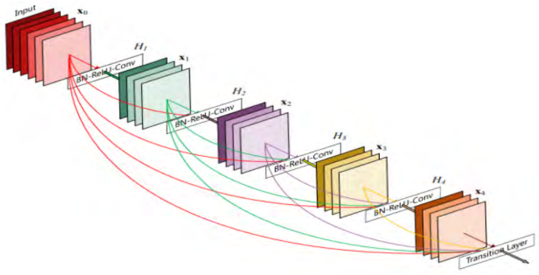

FIGURE 8. DenseNet with 5 layers [8].

B. LABEL BASED METRICS

This category exploits the use of two types of averaging method. The former is called macro-average, where first any binary evaluation metric can be computed on each individual class and then averaged over all classes, while the latter is called micro-average, where any binary evaluation metric can be computed globally over all instances and all classes [25]. Recently, the area under the receiver operating characteristic (ROC) curve, known as the AUC, has been widely used in MLC because it avoids the supposed subjectivity in the threshold selection process, when continuous probability derived scores are converted to a binary presence–absence variable, by summarizing overall model performance over all possible thresholds [26].

VII. RESULTS AND DISCUSSION

In this section, we conduct the experiments on two CXR datasets described in Section. 5.1. As shown in Table 2 each dataset was randomized and then split into 80% of the training set and 20% of the testing set. For performance evaluation, we used the following metrics AUC, hamming loss and

TABLE 2. Total number of training and testing CXRs per dataset.

Dataset	Train	Test
ChestX-ray14	89692	22424
CheXpert	107461	26866

TABLE 3. Results on ChestX-ray14 dataset.

Classifier	Training time	Metrics		
		Micro-F1	Hamming Loss	Average AUC
BR (LR)	21 min	0.547	0.061	**0.877**
LP (LR)	4h50min36s	0.540	**0.069**	0.875
CC (LR)	**17min9s**	**0.561**	0.067	0.876

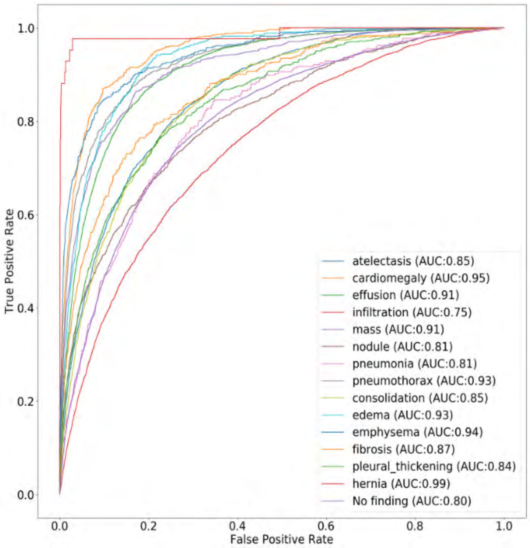

FIGURE 9. The AUC results and ROC curves obtained by BR on ChestX-ray14 dataset.

micro-averaged F1 score. All experiments are run on an HP ZBook17 with Intel Core i7-4700MQ CPU, 24 GB RAM, and NVIDIA Quadro K3100M with 4 GB.

A. RESULTS ON CHESTX-RAY14 DATASET

Table 3 presents a summary of classification performance results, in terms of micro-averaged F1 score, hamming loss and averaged AUC. As can be seen, Classifier Chain is the fastest classifier with a training time of 17min9s, followed by Binary Relevance with a training time of 21 min, while Label Powerset takes the longest time, up to 5 hours. The three classifiers produce very satisfactory and close results where Binary Relevance achieved the best hamming loss and averaged AUC of 0.061 and 0.877 respectively, while Classifier Chain reached the highest micro-averaged F1score of 0.561.

It is obvious from the above figures that our classifiers achieve very close and high AUC values, between 0.75 and 0.99, in all pathology labels of the ChestX-ray14 dataset.

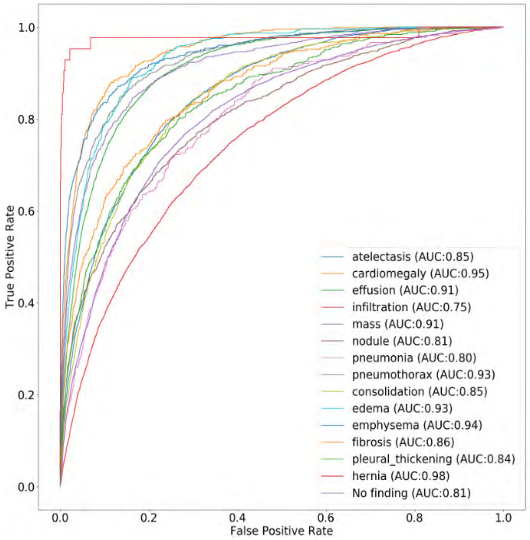

FIGURE 10. The AUC results and ROC curves obtained by LP on ChestX-ray14 dataset.

TABLE 4. Results on CheXpert dataset.

Classifier	Training time	Metrics		
		Micro-F1	Hamming Loss	Averaged AUC
BR(LR)	**18min47s**	0. 622	**0.116**	**0.812**
LP(LR)	7h15min53s	0.619	0.126	0,808
CC(LR)	22min33s	**0.631**	0.118	0.785

B. RESULTS ON CHEXPERT DATASET

As shown in Table 4, Binary Relevance takes the least training time on CheXpert dataset and achieves the best hamming loss of 0.116 and the highest averaged AUC of 0.812. Classifier Chain takes also little time in the training process and reaches the highest micro-F1 score of 0.63, while training the Label Powerset classifier on CheXpert dataset takes a long time of up to 7 hours. It should also be mentioned that the three classifiers give very similar results.

Figures 11-13 illustrate the AUC values and the ROC curves obtained by each classifier on the 14 pathology labels. It is clear that the performance of the three classifiers in terms of AUC values is almost similar and very high for all pathology labels.

Although the datasets suffer from the problem of imbalance label distribution (see Figure 3, 4, 6, and 7) that has been exacerbated by the use of problem transformation methods, especially for minority labels. Our method performs very well on ChestX-ray14 dataset, that contains only frontal CXRs, as well as on CheXpert dataset that provides frontal and lateral CXRs.

It is very likely that the presence of certain pathology could determine if another is also likely to be present or not. For this, we used CC and LP to exploit the correlation between pathologies and overcome the label independence assumption

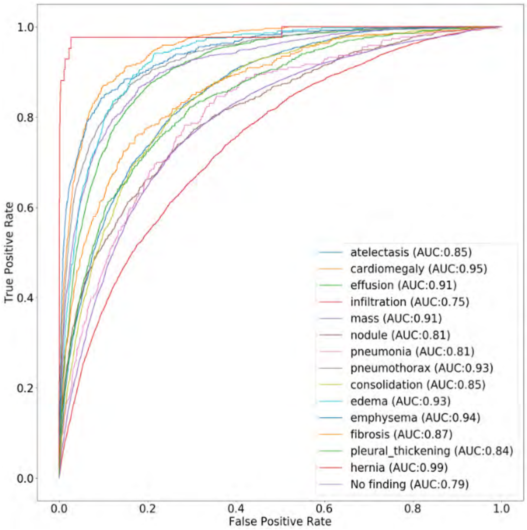

FIGURE 11. The AUC results and ROC curves obtained by CC on ChestX-ray14 dataset.

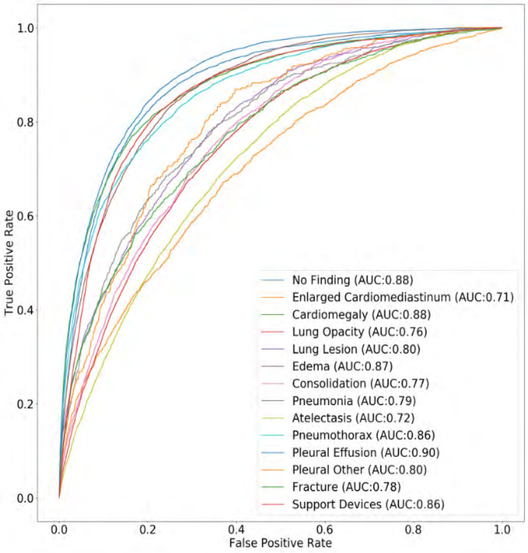

FIGURE 12. The AUC results and ROC curves obtained by BR on CheXpert dataset.

of BR. However, in our case, these methods did not give us the desired results because they did not provide any significant improvement compared to BR.

C. COMPARISON TO THE STATE-OF-THE-ART METHODS

Here, we compare the results obtained by our proposed method with the state-of-the-art results on the well-known chestX-ray14 dataset. To be fair in our comparison, we used the same train/test distribution as other methods with 70% for training, 20% for testing, and we ignored the remaining 10% because we did not need the validation process. As a result, we noticed that this change did not affect the performance results of our problem transformation methods.

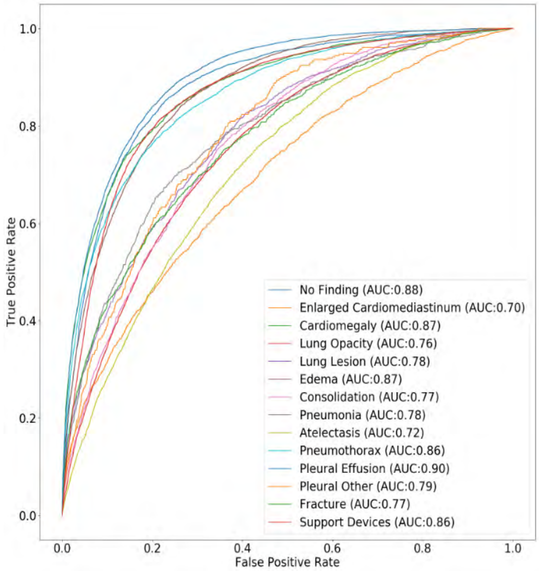

FIGURE 13. The AUC results and ROC curves obtained by LP on CheXpert dataset.

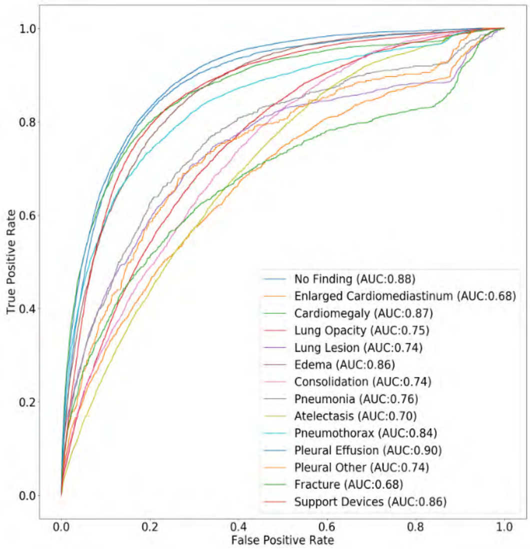

FIGURE 14. The AUC results and ROC curves obtained by CC on CheXpert dataset.

Our proposed method outperformed the current state-of-the-art by an averaged AUC of 1.1%. As can be seen in Table 5, the three proposed classifiers yield the best per-class AUC in 12 pathology labels including Atelectasis, Cardiomegaly, Infiltration, Mass, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural_thickening and Hernia, while the highest AUC of 0.986 among all pathology labels is attained on Hernia.

More importantly, this comparison proves the validity of the basic assumption of our study, namely that replacing the trainable classifier, the last fully connected layer, of a CNN model by the powerful multi-label problem transformation methods will improve the classification performance. The proof is that our method exceeds

TABLE 5. Comparison to the state-of-the-art methods on ChestX-ray14 dataset.

Pathology Label	CheXNet [23]	AG-CNN [24]	BR(LR)	LP(LR)	CC(LR)
Atelectasis	0.821	0.853	**0.855**	0.854	**0.855**
Cardiomegaly	0.905	0.939	**0.952**	0.948	**0.952**
Effusion	0.883	0.903	**0.911**	0.910	0.909
Infiltration	0.720	**0.754**	0.749	**0.748**	**0.747**
Mass	0.862	0.902	**0.913**	0.912	**0.913**
Nodule	0.777	**0.828**	0.810	0.809	0.810
Pneumonia	0.763	0.774	0.809	0.804	**0.811**
Pneumothorax	0.893	0.921	**0.929**	0.928	**0.929**
Consolidation	0.794	0.842	**0.849**	**0.849**	0.848
Edema	0.893	0.924	**0.931**	**0.931**	**0.931**
Emphysema	0.926	0.932	**0.942**	0.941	**0.942**
Fibrosis	0.804	0.864	**0.867**	0.861	0.865
Pleural_thickening	0.814	0.837	**0.848**	0.843	0.843
Hernia	0.939	0.921	**0.986**	0.977	**0.986**
Averaged AUC	0.842	0.871	**0.882**	0.880	0.881
No Finding	-----	------	0.804	**0.811**	0.794

DenseNet-121 model [23] results on all pathology classes by an average AUC of 4%.

VIII. CONCLUSION

In this paper, we propose a new approach that combines the effectiveness of CNN for image feature extraction and the power of supervised multi-label classifiers in order to tackle the task of thorax diseases detection on CXRs. The task has been carried out with a pre-trained DenseNet-121 model as feature extractor and different problem transformation methods such as BR, LP, and CC. The evaluation process was conducted using performance metrics like hamming loss, micro-averaged f1 score, and average AUC. The results showed that our method achieved great results and outperformed current state-of-the-art on ChestX-ray14 dataset. To further substantiate the results of this study, several improvements could be made, such as the use of an attention mechanism to improve CNN's work and train our classifier on a more balanced data set to avoid the problem of imbalance label distribution.

REFERENCES

[1] A. Brady, R. Ó. Laoide, P. McCarthy, and R. McDermott, "Discrepancy and error in radiology: Concepts, causes and consequences," *Ulster Med. J.*, vol. 81, no. 1, pp. 3–9, Jan. 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3609674/

[2] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Comput.*, vol. 29, no. 9, pp. 2352–2449, Sep. 2017. [Online]. Available: https://www.mitpressjournals.org/doi/10.1162/neco_a_00990

[3] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[4] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, E. Soria, J. Martin, R. Magdalena, M. Martinez, and A. Serrano, Eds. Hershey, PA, USA: IGI Global, 2009, ch. 11, p. 23.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14. [Online]. Available: https://arxiv.org/abs/1409.1556

[8] G. Huang, Z. Liu, K. Q. Weinberger, and L. Van der Maaten, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2016, pp. 4700–4708. [Online]. Available: https://arxiv.org/abs/1608.06993

[9] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognit.*, vol. 37, no. 9, pp. 1757–1771, Sep. 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320304001074

[10] G. Tsoumakas and I. Vlahavas, "Random k-labelsets: An ensemble method for multilabel classification," in *Proc. 18th Eur. Conf. Mach. Learn.*, 2007, pp. 406–417.

[11] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," in *Machine Learning and Knowledge Discovery in Databases*, vol. 5782. 2009, pp. 254–269. doi: 10.1007/s10994-011-5256-5.

[12] M.-L. Zhang and Z.-H. Zhou, "A k-nearest neighbor based algorithm for multi-label classification," in *Proc. IEEE Int. Conf. Granular Comput. (GrC)*, Jul. 2005, pp. 718–721.

[13] A. Clare and R. D. King, "Knowledge discovery in multi-label phenotype data," in *Proc. 5th Eur. Conf. PKDD*, 2001, pp. 42–53.

[14] A. Jiang, C. Wang, and Y. Zhu, "Calibrated rank-SVM for multi-label image categorization," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2008, pp. 1450–1455.

[15] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning," *Pattern Recognit.*, vol. 45, no. 9, pp. 3084–3104, Sep. 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320312001203

[16] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Random k-labelsets for multilabel classification," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 7, pp. 1079–1089, Jul. 2011.

[17] J. Read, "A pruned problem transformation method for multi-label classification," in *Proc. New Zealand Comput. Sci. Res. Student Conf.*, 2008, pp. 143–150.

[18] P. Szymański, T. Kajdanowicz, and K. Kersting, "How is a data-driven approach better than random choice in label space division for multi-label classification?" *Entropy*, vol. 18, no. 8, p. 282, Jun. 2016. doi: 10.3390/e18080282.

[19] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3462–3471.

[20] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, and K. Lyman. (2017). "Learning to diagnose from scratch by exploiting dependencies among labels." [Online]. Available: https://arxiv.org/abs/1710.10501

[21] P. Kumar, M. Grewal, and M. M. Srivastava, "Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs," 2017, *arXiv:1711.08760*. [Online]. Available: https://arxiv.org/abs/1711.08760

[22] P. Rajpurkar *et al.* (2017). "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning." [Online]. Available: https://arxiv.org/abs/1711.05225

[23] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang. (2018). "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification." [Online]. Available: https://arxiv.org/abs/1801.09927

[24] J. Rubin, D. Sanghavi, C. Zhao, K. Lee, A. Qadir, and M. Xu-Wilson. (2018). "Large scale automated reading of frontal and lateral chest X-rays using dual convolutional neural networks." [Online]. Available: https://arxiv.org/abs/1804.07839

[25] J. Irvin *et al.* (2019). "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison." [Online]. Available: https://arxiv.org/abs/1901.07031

[26] J. M. Lobo, A. Jiménez-Valverde, and R. Real, "AUC: A misleading measure of the performance of predictive distribution models," *Global Ecol. Biogeogr.*, vol. 17, no. 2, pp. 145–151, Mar. 2008. [Online]. Available: https://scinapse.io/papers/2141014056

IMANE ALLAOUZI received the Engineering degree in computer science from the National School of Applied Sciences, Al-Hoceima, Morocco, in 2014. She is currently pursuing the Ph.D. degree with the Faculty of Sciences and Techniques, Abdelmalek Essaadi University, Tangier, Morocco. Her research focuses on the application of artificial intelligence techniques in medicine and healthcare. She has authored six papers published in international conferences.

MOHAMED BEN AHMED received the Ph.D. degree in computer sciences and Telecommunications from Abdelmalek Essaadi University, in 2010. He is currently an Associate Professor of computer sciences with Abdelmalek Essaadi University, Morocco. He is also a Supervisor of several theses and an Investigator of several international research projects about smart cities. He has authored more than 20 papers published in international journals and conferences. His researches are about data mining, routing in wireless sensor networks, and smart cities.