# Deep learning to jointly analyze images and clinical data for disease detection

Federica Crobu, Agostino Di Ciaccio

**Abstract** In recent years, computer-assisted diagnostic systems increasingly gained interest through the use of deep learning techniques. Surely, the medical field could be one of the best environments in which the power of the AI algorithms can be tangible for everyone. Deep learning models can be useful to help radiologists elaborate fast and even more accurate diagnosis or accelerate the triage systems in hospitals. However, differently from other fields of works, the collaboration and co-work between data scientists and physicians is crucial in order to achieve better performances. With this work we show how it is possible to classify X-ray images through a multi-input neural network that also considers clinical data. Indeed, the use of clinical information together with the images allowed us to obtain better results than those already present in the literature on the same data.

**Key words:** Deep Learning, Medical Deep Learning, Convolutional Neural Networks, X-ray images, Multi-input Neural Networks

## 1 Introduction

Recent years have been marked by an exponential growth of interest towards whatever concerns data. Thanks to their great availability and hardware/software breakthroughs, many improvements and progresses have been made in the world of deep learning [5]. The use of deep convolutional neural networks has had a great impact on image recognition techniques. In this context, the evolution of research conducted in the last decade is well represented by the continuous progress in the

Federica Crobu
Sapienza Università di Roma, e-mail: federicacrobu@gmail.com

Agostino Di Ciaccio
Sapienza Università di Roma, e-mail: agostino.diciaccio@uniroma1.it

ImageNet dataset [9], until few years ago considered the benchmark for new architectures.

Among the many scopes of the AI, one of the most fascinating and advantageous branches for the application of these models is medicine. However the challenge is more complex: while within the general context of image recognition, the goal is to classify what it is contained in a given image (since the information is completely inside the picture), in the specific case of medical images we also should consider other important information about the patients. In fact, in order to try to emulate the role of an expert radiologist, the model should consider much more information such as demographic and clinical details.

Doctors usually gather and handle all this information and it is equally advantageous to provide them to the predictive model. From the technical point of view, the goal of including more inputs of different nature can be achieved using a multi-input neural network architecture. Using this kind of model we were able to obtain a very accurate classification, as shown in the following sections. Until a few years ago, it was unreasonable to think about a future in which doctors would be helped by computers to recognize diseases and elaborate diagnoses. The impact of these new technologies could represent a drastic improvement in underdeveloped countries, where the availability of doctors can often be problematic and pathologies such as pneumonia are still one of the main causes of death. Moreover, it could also be helpful in wealthy countries, where the number of radiologists is insufficient.

## 2 State-of-the-art and challenges of the medical deep learning

Among the many studies, some stand out for having achieved an accuracy comparable with that of the radiologists. DeepMind and Google Health have successfully trained an algorithm on mammogram images from a large database of 28,953 female patients in the US and UK, the results were published in the journal Nature [10]. Considering 2 images for each breast, they analysed 115,812 images. In a standard analysis, about 20% of screenings fail to find breast cancer even when it is present and many others are false positive. The AI algorithm decreased both types of error performing better than human radiologists (AUC 0.889 for UK data). To correctly evaluate the results, the real outcomes were derived from the biopsy record and longitudinal follow-up. NYU researchers published a similar study [11] using 229,426 screening mammography exams on 141,473 patients with about 1 million of images. Their network achieved an AUC of 0.895 and, to validate the model, they conducted a reader study with 14 radiologists, each analyzing 720 exams. The model confirmed its goodness showing an accuracy higher than a single experienced radiologist. However, both studies concluded that the AI screenings should be used in tandem with radiologists. In fact, thanks to the combination of experienced doctors and computers, it is possible to obtain the most precise diagnostic results.

The benefits of AI systems in automated triaging of chest radiographs have been explored by Annarumma et al. [2]. This work uses more than four hundred thou-

sands X-rays, jointly with their reports. Firstly, a NLP algorithm extracts the prioritization level from each report, subsequently a DCNN model associates the urgency from the image's analysis. The new prioritization system was tested in a simulation study, which showed a shorter mean delay for critical cases.

To apply these methodologies, an important requirement is the availability of a large and reliable database of images and clinical evaluations. This need clashes with the complexity of the image labeling process, as the definition of diagnosis is always characterized by a certain subjectivity, even if made by expert radiologists.

In general, an optimal solution to the problems faced in the medical area is to include much more information beyond the mere analysis of the images. For example, the correlation of certain pathologies to age or smoking is well known. Other diseases may be characterized by genetic predispositions and many diseases can be related to each other. Thus, the more additional information we have about the patients' clinical history the more we are able to construct a framework useful to improve the predictive model.

This work is based on our previous paper [4], a similar approach, given by Baltruschat [3], is discussed in section 3.

## 3 Material and methods

Among the general framework of medical deep learning we decided to focus on the analysis of X-ray images. Probably, the largest public database containing both images and clinical information is ChestX-ray14. From a technical point of view, we had to solve a multi-class and multi-label problem, since the task is the prediction of presence/absence of 14 diseases that can coexist in the same diagnostic image.

We will demonstrate how a multi-input neural network, so called since it is made by two independent nets joined in the end to perform predictions, can fruitfully use the information provided by the images with that coming from the patients' other data.

### *3.1 The data*

The *ChestX-ray14* [15] database was released in 2017 by the United States National Institutes of Health (NIH) and contains over 112,000 radiographic frontal chest images of 30,805 patients.

Each of them can be healthy or sick, affected by one or more of the following 14 diseases: Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural thickening, Pneumonia, Pneumothorax. Furthermore, a "no finding" category represents the images in which none of the previously mentioned diseases have been detected. "No finding" is the diagnosis in 60361 radiographs, while, for example, the diagnosis of "only" pneu-

monia is given in 322 cases. Few clinical and demographic information were also available: age, patient gender and follow-up number. In fact, each patient may have performed more than one radiographic examination, the progressive follow-up number indicates the sequence in which the examination was performed.

The labels, corresponding to the pathologies identified in each image, were extracted from radiological reports using natural language processing techniques with an accuracy that is declared by the authors over 90% [15]. Therefore, we cannot fully trust the labelling process and, furthermore, some researchers have raised many doubts about the correctness of the labels. Most of the criticism has been advanced by the radiologist Luke Oakden-Rayner [12] who, after observing the images, stated that many incorrect labels are present in the data and thus he could not say what the algorithm would be really able to understand and learn from such images. In addition, reports are mainly written in order to help other doctors, and the labels extracted by them can be different from the final diagnosis of the physicians.

**Fig. 1** Some images of the database ChestX-ray14

### *3.2 Previous works on the same data*

This dataset has been already used by many other researchers. Surely, the best-known work was made by a Stanford's team [13]. They proposed an architecture called CheXNet based on the usage of the DCNN architecture called DenseNet-121 [8]. This work represents, at present, the state-of-the-art results in terms of AUC scores. Other important works are Yao's et al. [16] and Wang's et al. [15]. The first one is mainly based on an architecture consisting of a DenseNet as an encoder and on a recurrent neural network as a decoder. Wang tries to apply some of the most famous CNN architectures (excluding DenseNet), achieving the best results with ResNet-50 [7]. However, there are numerous other papers that address this problem on the same or similar data using a deep neural network. For example, [1] proposed to apply a pretrained CNN as a feature extraction from the images and then, in sequence, a classification model. Another interesting work is that of Gündel [6], but these papers did not use additional clinical data. More interesting, from our point of view, is the Baltruschat's [3] work.

The last paper uses a multi-input neural network and includes the analysis of 3 variables: age, sex and view position. The architecture of the model was based on the ResNet-50 model applied to 448x448 px images. Because of this choice, which implies different input sizes than those expected by the model (224x224 px),

the authors added as first layer one Max Pooling to reduce the size of the images. The three variables were concatenated and directly linked to the output. In their work, some choices were introduced that impede a direct comparison with other applications in the literature. In particular, they did not use the 'official partition' (the benchmark train/test split proposed by the authors of the database). Although they experimented different architectures based on the ResNet model, the results obtained seem worse than those of the previous papers. They stressed the importance of including clinical data, but they did not consider the patient's medical history, which, to some degree, could be derived from the data. These aspects and the model architecture constitute the main elements of differentiation from our proposal.

### *3.3 The model*

Inspired by the the Stanford's work, we decided to enrich the model by including the few clinical and demographic information available with these images. To reach our goal we employed two independent networks that are joined at the end in order to share information before making predictions (a schematic drawing can be seen in figure 3).

The first and main branch consists of a Convolutional Neural Network, suitable to capture the essence of the X-rays. Among the many possibilities available, we decided to adopt the DenseNet-121 model, which is a CNN with 121 layers. The aspect that characterizes the architecture is the presence of 4 dense blocks, respectively with 6-12-24-16 layers inside. The blocks are connected by transition blocks each consisting of one convolutional layer and one pooling layer, which have the task of reducing the dimensionality (see figure 2).

The potential of this architecture lies in the usage of a deep structure characterized by many "short paths" between the layers that constitute the network itself [8]. This innovative mechanism lets the information pass directly from a layer to all the other ones, in a feed-forward fashion. This model has shown to be very efficient in terms of optimization, achieving top performances on benchmark datasets as ImageNet.

The second and innovative step is the building of the parallel network which processes the non-image characteristics. It considers age, opportunely rescaled using min-max normalization, sex and other 14 new dummy variables using the follow-up information. In fact, we constructed these new variables by recording patient information obtained in the previous pathological history, if present in the data. This branch of the network includes one input layer with 16 neurons and two hidden dense layers with 128 neurons activated by a ReLU function.

Finally, the two networks are concatenated and connected to the output layer consisting of 14 neurons with sigmoid activation function, whose task is to estimate the probability of the presence of each disease in the X-ray image.

The data was divided using the official benchmark partition proposed by [15] which consists of 80% for the training set and 20% for the test set. To make the

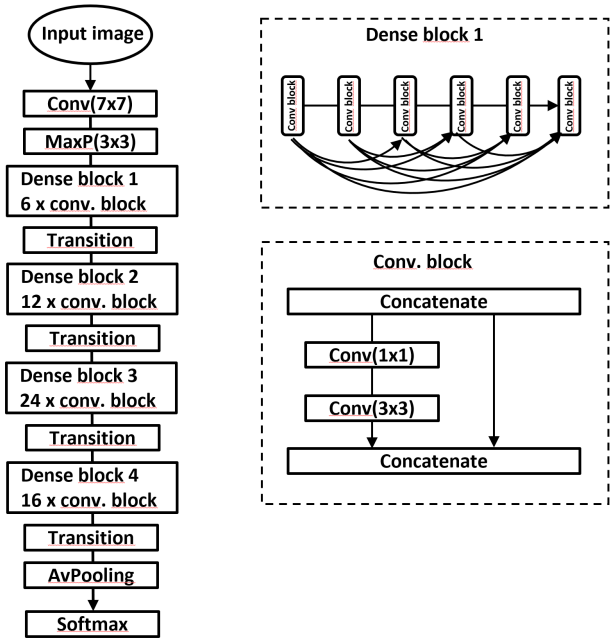


**Fig. 2** The DenseNet-121 architecture [8], based on the repetition of two kind of blocks: the *dense block*, able to perform the concatenation of many different convolution filters of different size, and the *transition block*, which performs the compression of the information. In order to make possible the last step, the CNN structure has to be flattened: this is performed using a Global Average Pooling layer.

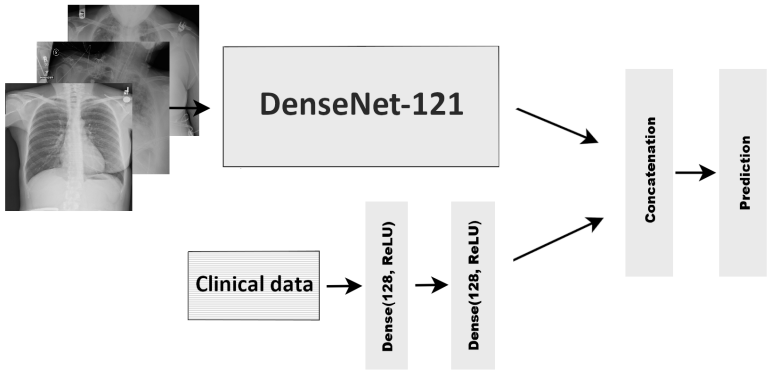


**Fig. 3** Multi-input neural network architecture. On the top the DenseNet-121 architecture [8] in which the 'top layers' have been eliminated. The branch at the bottom consists of two hidden dense layers applied to the non-image inputs. The two branches are then concatenated in order to produce predictions.

tuning of the model we used 20% of the training set as a validation set. The entire network has a complex structure with 123 'main' layers and more than 7 million parameters. We used the pretrained weights of DenseNet-121 (without the top-layers) on Imagenet as initialization of the convolutional neural network, while the second network has been trained from scratch using random weights. In the first epochs, to avoid the corruption of Imagenet's pre-trained weights, DenseNet's weights were frozen. To solve this multi-input multi-output problem, we have employed a weighted binary cross-entropy loss function [5] for accounting the high imbalance among the classes. Moreover, a data augmentation has been applied to the images. We tried several alternatives, for example adding noise or a slight zoom, but a simple horizontal flip of the X-ray resulted to be the best choice. As regard the optimization technique, we have chosen the *Adam* method with a tiny learning rate (0.001 and 0.0001 to fine tune). Adam combines the best properties of the AdaGrad and RMSProp algorithms to provide an optimization algorithm that can handle sparse gradients on noisy problems [5]. To perform the analysis we used the Tensorflow library and one Nvidia Titan XP 6100 GPU. To train one model it took up to 120 hours.

## 4 The results

Despite the limited clinical and demographic data available, our approach provides an interesting improvement of the state-of-the-art results, confirming our intuition of the architecture's power. Following the literature, we have adopted the AUC (Area Under the ROC Curve) index as the main tool to evaluate the quality of the predictions (figure 4). The table 1 shows the comparison of the performances of our model with the best results obtained by other researchers in terms of AUC scores.

It is evident in the table 1 that the average AUC has been significantly improved by our approach and, for most classes, we have clearly outperformed previous jobs. The scores show great variability: from 0.731 for Infiltrations to 0.966 for Hernia. The reason for these differences can be partly attributed to the imbalance of the data (even if we have applied appropriate weights to the training set), and partly to the differences between the pathologies: some of them are more difficult to identify with the available information.

## 5 Conclusions

The results of this application confirmed the validity of our approach: a multi-input neural network architecture can significantly improve predictions. Clearly, the idea of combining different heterogeneous sources of information can be applied in many other fields of medicine. Whenever the patient's clinical and/or demographic information is available, it is possible and fruitful to take this approach. Another possi-

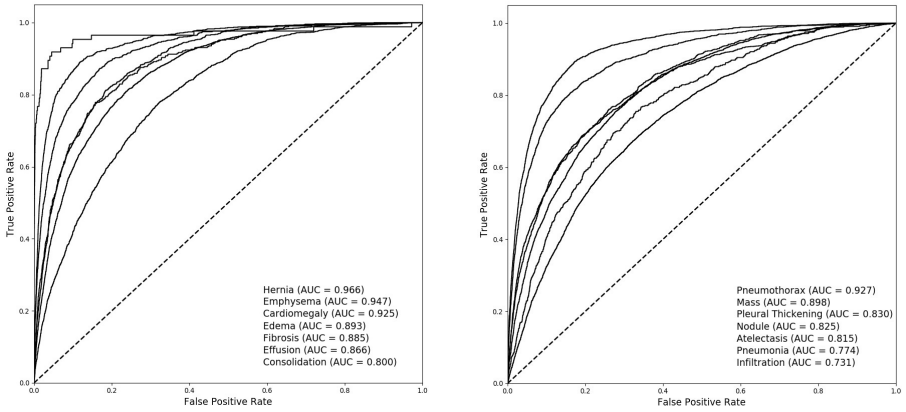


**Fig. 4** ROC curves of the 14 diseases on the training (left) and test (right). The diseases' curves are represented according to the decreasing AUC scores order.

**Table 1** AUC scores comparison

| | Wang et al. | Yao et al. | CheXNet | **Multi-input** |
|---|---|---|---|---|
| Official split | Yes | No | No | Yes |
| Atelectasis | 0.716 | 0.772 | 0.809 | **0.815** |
| Cardiomegaly | 0.807 | 0.904 | **0.925** | **0.925** |
| Effusion | 0.784 | 0.859 | 0.864 | **0.866** |
| Infiltration | 0.609 | 0.695 | **0.735** | 0.731 |
| Mass | 0.706 | 0.792 | 0.868 | **0.898** |
| Nodule | 0.671 | 0.717 | 0.780 | **0.825** |
| Pneumonia | 0.633 | 0.713 | 0.768 | **0.774** |
| Pneumothorax | 0.806 | 0.841 | 0.889 | **0.927** |
| Consolidation | 0.708 | 0.788 | 0.790 | **0.800** |
| Edema | 0.835 | 0.882 | 0.888 | **0.893** |
| Emphysema | 0.815 | 0.829 | 0.937 | **0.947** |
| Fibrosis | 0.769 | 0.767 | 0.805 | **0.885** |
| Pleural Thickening | 0.708 | 0.765 | 0.806 | **0.830** |
| Hernia | 0.767 | 0.914 | 0.916 | **0.966** |
| **Average** | 0.738 | 0.803 | 0.841 | **0.863** |

bility, which could produce great strides in medical AI, could be the joint real-time work with the radiologist [14]. In this way all the entities involved could enjoy significant advantages: doctors would be helped by the computer while analyzing the images and the algorithm would be trained in real-life situations, making a tangible contribution to its development.

Finally, it would be remarkable to have more public medical data in order to improve the researches, hoping that future studies in this sector will lead to a better quality of life and healthcare all over the world.

## References


1. Allaouzi, I., Ahmed, M.B.: A Novel Approach for Multi-Label Chest X-Ray Classification of Common Thorax Diseases. IEEE Access 7: 64279-64288 (2019)
2. Annarumma M., Withey S.J., Bakewell R.J., Pesce E., Goh V., Montana G.: Automated Triaging of Adult Chest Radiographs with Deep Artificial Neural Networks. Radiology 2018180921 (2019)
3. Baltruschat, I.M., Nickisch, H., Grass, M. et al.: Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification. Sci Rep 9, 6381 (2019) doi:10.1038/s41598-019-42294-8
4. Crobu F., Di Ciaccio A.: Classify X-ray images using convolutional neural networks. In: Porzio G. C., Greselin F., Balzano S.: CLADAG 2019 Book of short papers, pp. 136-139. Centro Editoriale di Ateneo Università di Cassino e del Lazio Meridionale, Cassino (2019)
5. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. The MIT Press (2016).
6. Gündel S., Grbic S., Georgescu B., Liu S., Maier A., Comaniciu D.: Learning to Recognize Abnormalities in Chest X-Rays with Location-Aware Dense Networks. In: Vera-Rodriguez et al. (eds) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, pp. 757-765. Springer, Cham (2019)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770-778 (2015)
8. Huang, G., Liu, Z., Weinberger, K.Q.: Densely Connected Convolutional Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261-2269 (2016)
9. ImageNet, Large Scale Visual Recognition Challenge (ILSVRC). http://image-net.org/challenges/LSVRC
10. McKinney, S.M., Sieniek, M., Godbole, V. et al.: International evaluation of an AI system for breast cancer screening. In: Nature 577, 89–94 (2020) doi:10.1038/s41586-019-1799-6
11. Wu N. et al.: Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. In: IEEE Transactions on Medical Imaging. doi: 10.1109/TMI.2019.2945514 (2019)
12. Oakden-Rayner, L.: Exploring the ChestXray14 dataset: problems (2017) https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/
13. Rajpurkar, P., et al.: CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv abs/1711.05225 (2017)
14. Wang P, Berzin TM, Glissen Brown JR, et al.: Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut 68:1813-1819 (2019)
15. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.: ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 3462–3471 (2017)
16. Yao, L., Poblenz, E., Dagunts, D., Covington, B., Bernard, D., Lyman, K.: Learning to diagnose from scratch by exploiting dependencies among labels. ArXiv abs/1710.10501 (2017)