Deep Transfer Learning for Interpretable Chest X-Ray Diagnosis

C. Lago, I. Lopez-Gazpio[(✉)], and E. Onieva

Faculty of Engineering, University of Deusto (UD),
Av. Universidades 24, 48007 Bilbao, Spain
carloslago@opendeusto.es, {inigo.lopezgazpio,enrique.onieva}@deusto.es

Abstract. This work presents an application of different deep learning related paradigms to the diagnosis of multiple chest pathologies. Within the article, the application of a well-known deep Convolutional Neural Network (*DenseNet*) is used and fine-tuned for different chest X-Ray medical diagnosis tasks. Different image augmentation methods are applied over the training images to improve the performance of the resulting model as well as the incorporation of an explainability layer to highlight zones of the X-Ray picture supporting the diagnosis. The model is finally deployed in a web server, which can be used to upload X-Ray images and get a real-time analysis.

The proposal demonstrates the possibilities of deep transfer learning and convolutional neural networks in the field of medicine, enabling fast and reliable diagnosis. The code is made publicly available (https://github.com/carloslago/IntelligentXray - for the model training, https://github.com/carloslago/IntelligentXray_Server - for the server demo).

Keywords: X-Ray diagnosis · Deep learning · Convolutional neural networks · Model interpretability · Transfer learning · Image classification

1 Introduction

Artificial Intelligence is poised to play an increasingly prominent role in medicine and healthcare due to advances in computing power, learning algorithms, and the availability of large datasets sourced from medical records and wearable health monitors [2]. In recent literature, deep learning shows promising results in medical specialities such as radiology [15], cancer detection [5], detection of referable diabetic retinopathy, age-related macular degeneration and glaucoma [19], and cardiology, in a wide range of problems involving cardiovascular diseases, performing diagnosis, predictions and helping in interventions [3].

In this article, we investigate the application of deep learning models for multiple chest pathology diagnoses with the objective of designing a fast and reliable method for diagnosing various pathologies by analyzing X-Ray images. For the

task, we employ, tune, and train a deep learning model using Tensorflow [1], as well as evaluate it on various medical state-of-the-art benchmarks through transfer learning. The used base model, DenseNet [11], is a state-of-the-art deep Convolutional Neural Network (CNN) that employs an innovative strategy to ensure maximum information flow between all the layers of the network. In the model, each layer is connected to all the others within a dense block, as a consequence, all layers can get feature maps from preceding ones. The resulting model is shown to be compact and have a low probability of overfitting [11]. Through this work, DenseNet is adapted to be evaluated in different benchmarks from the medical domain. A post-process training step based on image augmentation is also incorporated in order to increase its accuracy. Our contribution is two-fold. We first adapt DenseNet for chest X-Ray effective diagnosis and then the addition of the explainability layer. The evaluation of the new model is performed on X-Ray classification benchmarks, including (i) pneumonia detection task; (ii) detection of different pathologies which can be evaluated by doctors; (iii) a Covid-ChestXray detection dataset; which consists of an open dataset of chest X-Ray images of patients who are positive (or suspected) of COVID-19 or other viral and bacterial pneumonia.

This article is structured as follows. Section 2 presents the data used for building and evaluating models, in this case, three different medical image databases. Section 3 introduces the original model, as well as the modification proposed. Section 4 presents the inclusion of the explainability layer to the model, in order to make it capable of explaining its diagnoses. Section 5 analyses results obtained. Finally, Sect. 6 presents conclusions and future works.
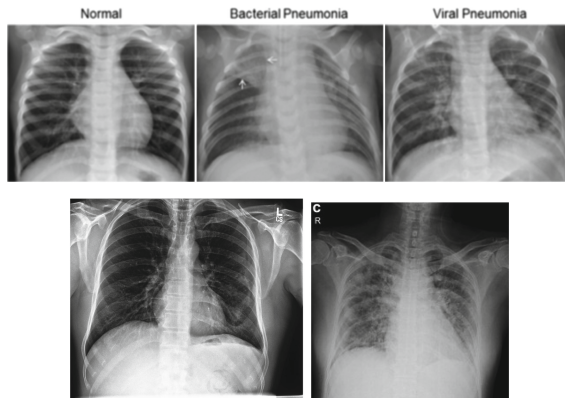
Fig. 1. (Top) Three examples from the pneumonia detection task [14]. (Bottom left) An example image from CheXpert interpretation task [12]. (Bottom right) An example from the Covid-ChestXray benchmark [6].

2 Data Description

This section describes the datasets and benchmarks used to fine-tune, train and evaluate the model. We also describe the preprocessing performed on the original data. The evaluation of the model is carried out in three different fields of the area of X-Ray diagnosis:

(i) *Chest X-Ray Images (Pneumonia)*, an open dataset of chest X-Ray images of patients and their pneumonia diagnosis [14].

(ii) *CheXpert*, a large dataset of chest X-Rays for automated interpretation, featuring uncertainty labels and radiologist-labeled reference standard evaluation sets [12].

(iii) *Covid-ChestXray-Dataset*, an open dataset of chest X-Ray images of patients who are positive (or suspected) of COVID-19 or other viral and bacterial pneumonias [6].

Data for the Pneumonia detection benchmark dataset (i) is publicly available from a Kaggle competition[1], including 5863 X-Ray images associated with two different resulting categories (Pneumonia in 25% of the cases and Normal in the remaining 75%). Automated chest X-Ray sources (ii) have been extracted from CheXpert database[2]. CheXpert is a dataset provided by Stanford ML Group, with over 224,316 samples with both frontal and lateral X-Rays collected from tests performed between October 2002 and July 2017 at Stanford Hospital. Images are labeled to differentiate 14 different diagnoses: no concluding pathology, presence of support devices, and a list of 12 different possible pathologies. The distribution of the classes is outlined in Fig. 2. Chest radiography is one of the most common imaging examinations performed overall, and it is critical for screening, diagnosis, and management of many life-threatening diseases. Automated chest radiography explainability is capable of providing a substantial benefit for radiologists. This research aims to advance the state-of-the-art development and evaluation of medical artificial intelligence applications. Finally, Covid-ChestXray-Dataset sources (iii) can be found at GitHub[3]. It contains around 100 X-Ray samples with suspected COVID-19. Data has been collected from public sources as well as through indirect collection from hospitals and physicians during the previous year. This task aims to differentiate between no pathology, pneumonia, and COVID-19 cases. Examples of X-Ray images used in the experimentation are presented in Fig. 1.

2.1 Data Preprocessing

To prepare the data for training, we first re-scaled all the input image sizes to 224×224 pixels, since not all the images from all the sources have the same resolution. In addition, image augmentation techniques increase the training set

---

[1] https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia.
[2] https://stanfordmlgroup.github.io/competitions/chexpert/.
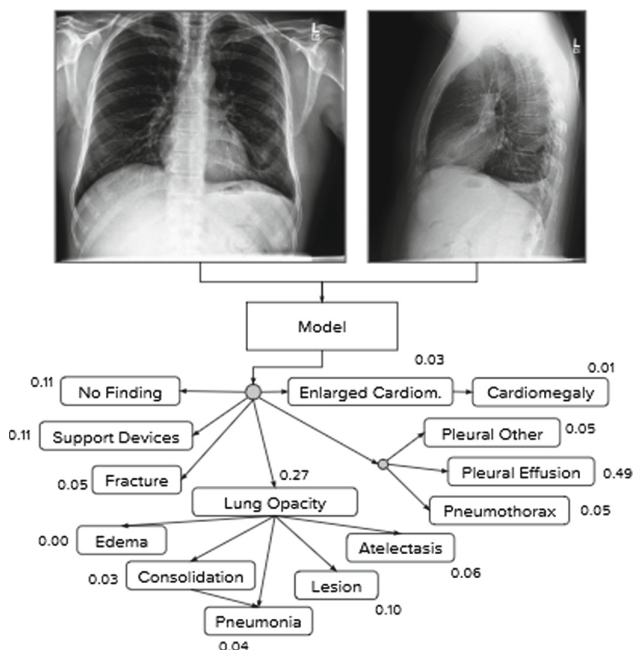[3] https://github.com/ieee8023/covid-chestxray-dataset.

Fig. 2. CheXpert classes distribution [12]

size to avoid overfitting [22]. The CheXpert dataset contains four possible labels, *empty* when the pathology was not considered (they are considered negative), *zero*, denoting the pathology is not present, *one* when the pathology is detected, and a value (-1) denoting that it is unclear if the pathology exists or not. For modeling, all the cases with empty or zero values are considered negative cases. In the case of unclear values, different approaches are applied depending on the pathology, according to the best results showed in [12] when using the U-zeros or U-ones method. Train, validation, and test splits are shown in Table 1 for all the different cases. Both the Pneumonia and the CheXpert cases have a split of 80/10/10, while the COVID-19 case has a split of 60/15/25, as there is less data available.
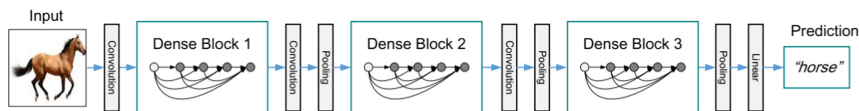

Fig. 3. DenseNet architecture overview. (Figure from [10]).

Table 1. Data balance in training and testing splits

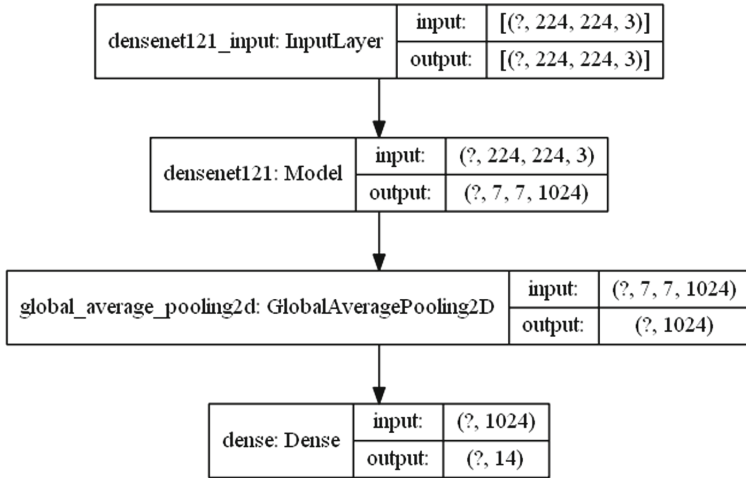| Case	Pathologies	Train and validation		Test	
		Positive	Negative	Positive	Negative
Pneumonia	Pneumonia	0.63	0.37	0.74	0.26
COVID-19	Pneumonia	0.31	0.69	0.34	0.66
	COVID-19	0.29	0.71	0.33	0.67
CheXpert	Cardiomegaly	0.64	0.36	0.47	0.53
	Edema	0.48	0.52	0.29	0.71
	Consolidation	0.21	0.79	0.25	0.75
	Atelectasis	0.63	0.37	0.59	0.41
	Pleural Effusion	0.53	0.47	0.38	0.62

3   Model Construction

CNNs have become the dominant Machine Learning approach for visual object recognition in recent years [10]. They perform convolution instead of matrix multiplication in contrast to fully connected neural networks. As a consequence, the number of weights is decreased resulting in a less complex network that is easier to train [21]. Furthermore, images can be directly imported into the input layer of the network avoiding the feature extraction procedure widely used in more traditional machine learning applications. It should be noted that CNNs are the first truly successful deep learning architectures due to the inherent hierarchical structure of the inner layers [10,17].

Deep CNNs can represent functions with high complexity decision boundaries as the number of neurons in each layer or the number of layers is increased. Given enough labeled training instances and suitable models, Deep CNN approaches can help establish complex mapping functions for operation convenience [17].

This research is based on the well-known DenseNet model [25], which is a popular architecture making use of deep CNNs. The main contribution of DenseNet relies upon that it connects all layers in the network directly with each other, and in that each layer also benefits from feature maps from all previous layers [10]. A visual representation of DenseNet is provided in Fig. 3. In this sense, DenseNet provides shortcut connections through the network that lead to deep implicit supervision, which is denoted in the state-of-the-art as a simple strategy to ensure maximum information flow between layers. This architecture has been used in a wide variety of benchmarks yielding state-of-the-art results as it produces consistent improvements in the accuracy, by increasing the complexity of layers, without showing signs for overfitting [25].
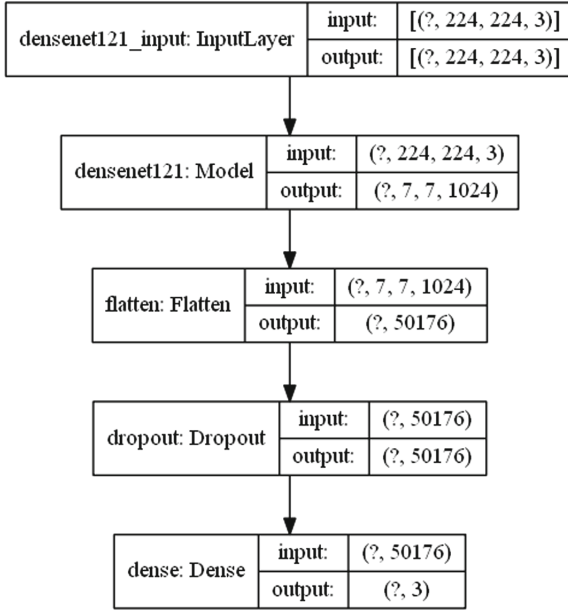
3.1   Fine-Tunning over DenseNet

Originally, DenseNet was trained for object recognition benchmark tasks as CIFAR-10, CIFAR-100, The Street View House Numbers (SVHN) Dataset or

| densenet121_input: InputLayer | input: | [(?, 224, 224, 3)] |
| | output: | [(?, 224, 224, 3)] |

| densenet121: Model | input: | (?, 224, 224, 3) |
| | output: | (?, 7, 7, 1024) |

| global_average_pooling2d: GlobalAveragePooling2D | input: | (?, 7, 7, 1024) |
| | output: | (?, 1024) |

| dense: Dense | input: | (?, 1024) |
| | output: | (?, 14) |

Fig. 4. DenseNet model architecture used for experiments (i) and (ii).

ImageNet. These benchmarks are composed of a big number of training instances for each one of the classes to predict [8,16,18]. The need for alteration of the original model raises when adapting it to environments where the number of training samples is low, as is the case of medical diagnosis based on an image. In order to accommodate the DenseNet network into the medical domain and reuse the latent knowledge condensed in its layers, the final layers of the network are updated to perform classification in the medical domain by retraining only these final layers. Transfer Learning is a great tool for fixing the lack of data problem widely extended in deep learning, which usually employs methods that need more data than traditional machine learning approaches. Transferring the knowledge from the source domain where the network has been trained to another is a common practice in deep learning [23]. Moreover, thanks to the transfer of knowledge, the amount of time it is required to learn a new task is decreased notably, and the final performance that can be achieved is potentially higher than without the transfer of knowledge [24]. All of our DenseNet model variants have been trained with transfer learning for DenseNet121, the only difference between each architecture is the regularization techniques that are implemented. As shown in Fig. 3, DenseNet includes several convolution layers (referred to as transition layers) and dense blocks in an iterative way, connecting all layers in the network directly with each other, with each layer receiving feature maps from all the previous layers [10].

The architectures of the base DenseNet models are showed in Figs. 4 and 5. The final Dense layer from the mentioned figures is the part that we fine-tune for experimentation. It should be noted that L2 regularization is being used on the GlobalAveragePooling2D layer, with a rate of 0.001, which is essential to avoid over-fitting in the network. This difference among architectures was found to produce better results in laboratory experimentation due to the lower sizes

densenet121_input: InputLayer	input:	[(?, 224, 224, 3)]
	output:	[(?, 224, 224, 3)]

densenet121: Model	input:	(?, 224, 224, 3)
	output:	(?, 7, 7, 1024)

flatten: Flatten	input:	(?, 7, 7, 1024)
	output:	(?, 50176)

dropout: Dropout	input:	(?, 50176)
	output:	(?, 50176)

dense: Dense	input:	(?, 50176)
	output:	(?, 3)

Fig. 5. DenseNet model architecture used for experiment (iii).

of the dataset. For both the first and second evaluation benchmarks, regarding the Pneumonia detection and the CheXpert identification, the same type of regularization techniques have been used. For training both models the following image augmentation techniques have been used: horizontal flip, zoom range both in and out in with a maximum value ($zoom_{range}$) of 0.2, ZCA whitening, to obtain decorrelated features, and rotation within a range (rot_{range}) of -5^a and $5°$. Furthermore, given the lower number of samples in the COVID-19 case, additional data augmentation was included: Height and width shift range, both with a range ($shift_{range}$) of 0.2, moving the image up to that percent vertically or horizontally, shear range ($shear_{range}$) of 0.2, rotation range of $5°$ and, brightness modification, ranging from 0.8 to 1.2. In addition, lateral images are being used for training the model too, which is used in this research as a form of image augmentation, which gives to the model a different viewpoint of a particular sample. Regarding the COVID-19 case, which also contains sample instances from Pneumonia and Covid cases, the network is slightly different as it uses dropout, which was observed useful not to over-train the model as several neuron units are randomly disconnected at training [9]. The specific model for this third experiment is shown in Fig. 5. It should be noted that for this case more image augmentation techniques are being used: horizontal flips, zoom range, rotation range, height and width shift range, shear range, and brightness range.

4   Explainability Layer

As a final step, an explainability layer is added on top of the adapted DenseNet to make the model self-explainable. Developing the explainability of a machine learning model is essential to understand the decision process behind the predictions of a model, analyzing if it makes sense or not, and facilitate human trust in the final decisions. It can also help to gain some insights and confidence in the model, often seen as a black-box tool, by observing clearly how it performs under given circumstances. Adding interpretability to models is a key towards transforming an untrustworthy model into a trustworthy one [20]. Local Interpretable Model-agnostic Explanations (LIME) [20] is used to provide explainability to the final model. LIME is an algorithm able to explain the predictions of a regressor or classifier by approximating its result with an interpretable model. LIME provides local explanations of predictions of a model by fitting an explanation model locally around the data point of which classification is to be explained. LIME supports generating model explanations for text and image classifiers. The layer implements the function in Eq. 1.

$$\xi(x) = g \in G \ argmin_x \ \mathcal{L}\left(f, g, \pi_x\right) + \Omega(g), \tag{1}$$

where the fidelity function $\mathcal{L}$ is a measure of how unfaithful an explanation g, an element in the set of possible interpretation models G, in approximating f, the probability of x belonging to a class in the locality defined by the proximity measure π_x. The $\Omega(g)$ term is a measure of complexity of the explanation of $g \in G$. For the explanation to ensure interpretability and local fidelity, it is necessary to minimize $\mathcal{L}\left(f, g, \pi_x\right)$ and have a $\Omega(g)$ low enough so it is interpretable by humans. *SP-LIME* is a method that selects a set of instances with explanations that are representative to address the problem of trusting the model. To understand how the classifier works, it is necessary to explain different instances rather than just provide an explanation of a single prediction. This method selects some explanations that are insightful and diverse to represent the whole model.

5   Experimentation and Results

This section presents results of the experimentation for each of the studied benchmark tests. For all cases, the performance of the model is evaluated in terms of Accuracy (Eq. 2), AUC (Eq. 3)[4] and micro and macro F1 scores (Eqs. 4 and 5 respectively).

$$Accuracy = \frac{TrueNegatives + TruePositive}{All \ samples} \tag{2}$$

$$AUC(f) = \frac{\sum_{t_0 \in \mathcal{D}^0} \sum_{t_1 \in \mathcal{D}^1} \mathbf{1}\left[f\left(t_0\right) < f\left(t_1\right)\right]}{|\mathcal{D}^0| \cdot |\mathcal{D}^1|} \tag{3}$$

---

[4] $\mathbf{1}\left[f\left(t_0\right) < f\left(t_1\right)\right]$ returns 1 if $f\left(t_0\right) < f\left(t_1\right)$ and otherwise 0, $\mathcal{D}^1$ is the set of positive examples and $\mathcal{D}^0$ of negatives.

Table 2. Pneumonia benchmark experimentation results.

Case	Accuracy	AUC	Macro-F1	Micro-F1
Training set	94.5%	0.99	0.96	0.96
Test set	93.75%	0.99	0.96	0.96

Table 3. CheXpert benchmark experimentation results.

Case	Accuracy	AUC	Macro-F1	Micro-F1
Training set	88%	0.93	0.88	0.9
Test set	83%	0.88	0.74	0.78

Table 4. Detailed CheXpert benchmark experimentation results.

Case	Accuracy	AUC	Macro-F1	Micro-F1
Cardiomegaly	74.3%	0.74	0.74	0.74
Edema	82.64%	0.78	0.78	0.83
Consolidation	77.78%	0.65	0.66	0.78
Atelectasis	70.83%	0.68	0.68	0.71
Pleural Effusion	82.41%	0.83	0.82	0.82

Table 5. COVID-19 benchmark experimentation results.

Case	Accuracy	AUC	Macro-F1	Micro-F1
Training set	92%	0.98	0.92	0.93
Test set	94%	0.98	0.94	0.94

$$F1_{micro} = 2 \times \frac{recall_{micro} \times precision_{micro}}{recall_{micro} + precision_{micro}} \tag{4}$$

$$F1_{macro} = \sum_{classes} \frac{F1\ of\ class}{number\ of\ classes} \tag{5}$$

For the calculation of micro F1, the precision and recall values are obtained with Eqs. 6 and 7, as well as for the calculation of macro F1 (Eq. 5). The calculation per class of the F1 values is done as presented in Eq. 8.

$$precision_{micro} = \frac{\sum_{classes} TP\ of\ class}{\sum_{classes} TP\ of\ class + FP\ of\ class} \tag{6}$$

$$recall_{micro} = \frac{\sum_{classes} TP\ of\ class}{\sum_{classes} TP\ of\ class + FN\ of\ class} \tag{7}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{8}$$

Pneumonia detection task results can be seen in Table 2 and the COVID-19 case results are shown in Table 5. We can see that the network has good performance on balanced datasets, proved by the macro and micro F1 results. CheXpert benchmark results can be seen in Table 3, with the results for the most important pathologies in detail in Table 4. Concerning CheXpert, it is much more complex, having 14 different observations and unknown labels. Moreover, only a small subset of the data has been used, as the objective is to prove the efficiency of this type of network but not get the best possible result. The F1 score only takes into account Cardiomegaly, Edema, Consolidation, Atelectasis, and Pleural Effusion, due to the unbalanced samples in the dataset. Model evaluation in benchmarks (i) and (iii) yields an accuracy of over 90%, with AUC scores of 0.98+, and lower results for the benchmark (ii), with accuracies over 83% and AUC scores of 0.9. These results are in line with other neural network approaches applied to chest X-Ray benchmarks. For instance, Çallı et al. in [4] report a total of 9 neural network systems trained and validated in the COVID-19 infection benchmark. Their results indicate that the mean accuracy of these systems is 0.902 with a standard deviation of 0.044. Our experimentation results also confirm that transfer learning can successfully be applied for rapid chest X-Ray diagnosis and help expert radiologists with a system that offers immediate assistance, as supported by [4].

5.1   Evaluating Model Explainability

With the addition of an explainability layer on top of it, the model can explain its behavior, by highlighting the areas of the image that support its diagnosis. The addition of this explanatory layer is the first step towards increasing human confidence in health-related artificial intelligence applications, as the model can self-explain its decisions to gain trust in human-computer interactions. We now show model explanations for each one of the benchmark datasets in which we evaluate our model. First, we show a sample instance from the pneumonia detection dataset (Fig. 6) The figure represents a pneumonia detectable by the airspace consolidation that can be seen on the right lower zone . The model correctly outlines the area affected by pneumonia. For the CheXpert benchmark dataset, the case shown in Fig. 7 has been diagnosed with cardiomegaly . Cardiomegaly is present when the heart is enlarged, as can be easily seen on the frontal X-Ray from Fig. 7. In this scenario, the network successfully performs a correct prediction of the area. Finally, we show an example for the COVID-19 benchmark dataset in Fig. 8 which has been diagnosed with airspace opacities . The model explanations show that the network is focusing on the relevant areas related to these opacities.
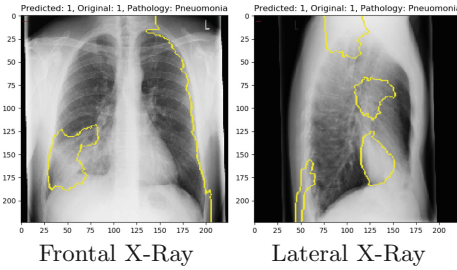
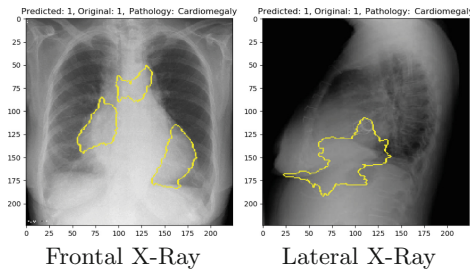**Fig. 6.** Pneumonia detection samples including model explanation.

Fig. 7. CheXpert samples diagnosed with cardiomegaly including model explanation.
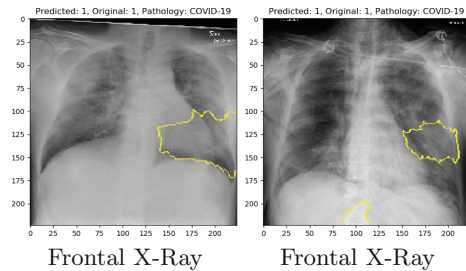

Fig. 8. COVID-19 samples diagnosed with COVID-19 including model explanation.

5.2   Demo

In addition to the development of the models, we build a web prototype where the CheXpert model is deployed. This web prototype enables uploading personal X-Ray images to get a diagnose in less than one second and model explanations in less than ten seconds. Figure 9 shows the web application for performing the predictions and how the results are presented. As shown in the image, it is possible to upload both the frontal and lateral X-Ray or just one for the prediction, the table on the right displays the prediction of the model for each of the pathologies along with its confidence in the prediction.
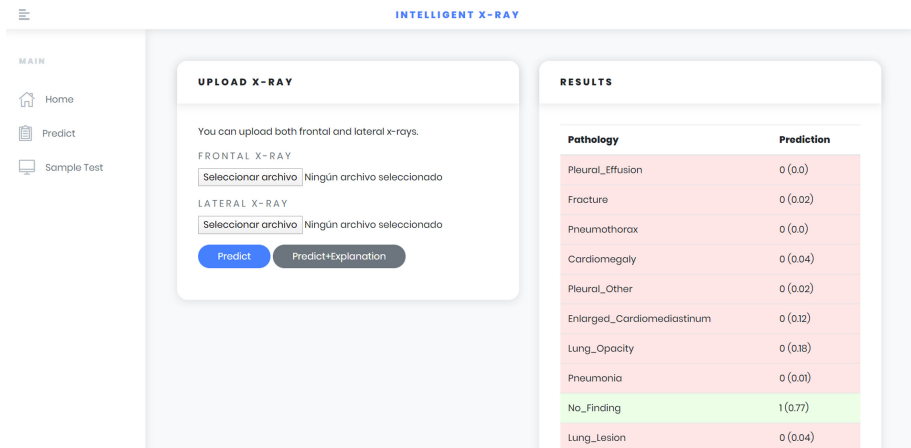
Fig. 9. Web demo

6 Conclusion and Future Work

This research demonstrates how deep learning and CNNs can be useful in the field of medicine, enabling fast and reliable support for diagnosis. For diseases like cancer, where early diagnoses could save millions of lives, they also enable a more accurate and early diagnosis, with models such as [13] achieving an accuracy of more than 90%. Furthermore, the addition of the explainability layer is an important step towards improving trust in model predictions, which is one of the main concerns for public usage. We consider the addition of this explanatory layer as the first step towards increasing human confidence in health-related artificial intelligence applications, as the model can self-explain its decisions.

Although polymerase chain reaction (RT-PCR) is the preferred way to detect COVID-19, the costs and response time involved in the process have resulted in a growth of rapid infection detection techniques, most of them based on chest X-Ray diagnosis [7]. The primary advantage of the automatic analysis of chest X-Rays through deep learning is that the technique is capable of accelerating the time required for the analysis. It should be noted that this study is just an experiment to showcase the capabilities of deep convolutional neural networks in the field of radiology, but shouldn't be considered as a way to replace a radiologist checkup. For future lines, it will be interesting to mix the current datasets used in the project with other existing datasets, enabling the detection of more pathologies and fixing the unbalance problems. More complex architectures could be tested in a bigger sample to improve performance.

References

1. Abadi, M., et al.: TensorFlow: a system for large-scale machine learning. In: 12th USENIX Symposium on Operating Systems Design and Implementation, pp. 265–283 (2016)

2. Ahuja, A.: The impact of artificial intelligence in medicine on the future role of the physician. PeerJ **7**, e7702 (2019)

3. Bizopoulos, P., Koutsouris, D.: Deep learning in cardiology. IEEE Rev. Biomed. Eng. **12**, 168–193 (2019). https://doi.org/10.1109/RBME.2018.2885714

4. Çallı, E., Sogancioglu, E., van Ginneken, B., van Leeuwen, K.G., Murphy, K.: Deep learning for chest x-ray analysis: a survey. Med. Image Anal. **72**, 102125 (2021)

5. Campanella, G., et al.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. **25**(1), 1301–1309 (2019). https://doi.org/10.1038/s41591-019-0508-1

6. Cohen, J.P., Morrison, P., Dao, L., Roth, K., Duong, T.Q., Ghassemi, M.: COVID-19 image data collection: prospective predictions are the future. arXiv preprint arXiv:2006.11988 (2020)

7. Das, N.N., Kumar, N., Kaur, M., Kumar, V., Singh, D.: Automated deep transfer learning-based approach for detection of COVID-19 infection in chest X-rays. IRBM (2020). https://www.sciencedirect.com/science/article/abs/pii/S1959031820301172

8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)

9. Gal, Y., Ghahramani, Z.: Bayesian convolutional neural networks with Bernoulli approximate variational inference. arXiv preprint arXiv:1506.02158 (2015)

10. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)

11. Iandola, F., Moskewicz, M., Karayev, S., Girshick, R., Darrell, T., Keutzer, K.: DenseNet: implementing efficient convnet descriptor pyramids. arXiv preprint arXiv:1404.1869 (2014)

12. Irvin, J., et al.: CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 590–597 (2019)

13. Kalra, S., et al.: Pan-cancer diagnostic consensus through searching archival histopathology images using artificial intelligence. NPJ Digit. Med. **3**(1), 1–15 (2020)

14. Kermany, D.S., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell **172**(5), 1122–1131 (2018)

15. Kim, J.R., et al.: Computerized bone age estimation using deep learning based program: evaluation of the accuracy and efficiency. AJR Am. J. Roentgenol. **209**(6), 1374–1380 (2017)

16. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Technical report, University of Toronto (2009)

17. Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., Alsaadi, F.E.: A survey of deep neural network architectures and their applications. Neurocomputing **234**, 11–26 (2017). https://doi.org/10.1016/j.neucom.2016.12.038

18. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. Technical report, NIPS Workshop (2011)

19. Rahimy, E.: Deep learning applications in ophthalmology. Curr. Opin. Ophthalmol. **29**(3), 254–260 (2018). https://doi.org/10.1097/ICU.0000000000000470
20. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016, pp. 1135–1144 (2016)
21. Ronao, C.A., Cho, S.-B.: Deep convolutional neural networks for human activity recognition with smartphone sensors. In: Arik, S., Huang, T., Lai, W.K., Liu, Q. (eds.) ICONIP 2015. LNCS, vol. 9492, pp. 46–53. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-26561-2_6
22. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. J. Big Data **6**(1), 1–48 (2019). https://doi.org/10.1186/s40537-019-0197-0
23. Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C.: A survey on deep transfer learning. In: Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I. (eds.) ICANN 2018. LNCS, vol. 11141, pp. 270–279. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01424-7_27
24. Torrey, L., Shavlik, J.: Transfer learning. In: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, pp. 242–264. IGI Global (2010)
25. Zhu, Y., Newsam, S.: DenseNet for dense flow. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 790–794 (2017)