International Journal of Computer Vision and Image Processing
Volume 12 • Issue 1 • January-March 2022

Screening of Radiological Images Suspected of Containing Lung Nodules

Raúl Pedro Aceñero Eixarch, Jaume I University, Spain
Raúl Díaz-Usechi Laplaza, Hospital General de Castellón, Spain
Rafael Berlanga Llavori, Jaume I University, Spain

ABSTRACT

This paper presents a study about screening large radiological image streams produced in hospitals for earlier detection of lung nodules. Being one of the most difficult classification tasks in the literature, the objective is to measure how well state-of-the-art classifiers can screen out the images stream to keep as many positive cases as possible in an output stream to be inspected by clinicians. The authors performed several experiments with different image resolutions and training datasets from different sources, always taking ResNet-152 as the base neural network. Results over existing datasets show that, contrary to other diseases like pneumonia, detecting nodules is a hard task when using only radiographies. Indeed, final diagnosis by clinicians is usually performed with much more precise images like computed tomographies.

KEYWORDS

Deep Learning, Image Classification, Patient Screening, Radiology

1. INTRODUCTION

Radiodiagnosis disease detection is a low-cost and universally widespread method. Its main drawback is that it must be carried out by highly qualified people (a radiologist) which are scarce in the public health system. Thus, during high workload in a hospital, most X-ray images go directly to the doctor without being reviewed by a radiologist. For this reason, it is necessary to create automatic screening tools able to redirect suspicious cases to radiologists so that some diseases can be detected at earlier stages. In this paper, we will focus on screening lung nodules associated to neoformative processes of lung cancer.

At present, the simple chest radiograph continues to be the most performed radiological examination on a daily basis, being interpreted by numerous specialists. Among them, radiologists are trained specifically for the interpretation of such images, but they represent a small percentage of all the specialties between which medicine is divided, so many of these studies are not reported by a radiologist.

The vast majority of pathologies express noticeable radiological findings, easily detectable by any doctor, and who can consult the image with a radiologist. But the initial stage of the neoformative process of lung cancer, before becoming a lung mass, is represented on the chest X-ray as a pulmonary nodule, in most cases in a very subtle way. Even trained eyes can miss these findings, and it is not uncommon that once the pathology is diagnosed, by reviewing previous examinations, we can

DOI: 10.4018/IJCVIP.20220101.oa1

appreciate or imagine clues. Lung tumors present different cell lines, and present in different ways in the lung parenchyma. They may have a central arrangement, being located in perihilar or even endobronchial topographies, as well as a more peripheral distribution. Initially, they will be represented as small pulmonary nodules, always smaller than 3cm, and once that figure is exceeded, they will become pulmonary masses. Densotomographically, they can present densities similar to those of soft tissue tissues, presenting an enhancement after the administration of intravenous contrast (due to their greater vascularization and neoangiogenesis). But this representation can vary, since they can present a necrotic center, hemorrhagic foci or even intratumoral cystic components. They generally have poorly defined borders, with a poor definition with respect to the adjacent structures, adopting a morphology of "spiculate borders", which as the disease progresses, can condition the invasion of organs and structures adjacent to the tumor. Given its variability in morphologies and locations where the pulmonary nodule can settle, it is important to recognize its presentation patterns both to identify them and to differentiate them from other pathologies (metastases, abscesses, interstitial diseases ... and much more).

For this reason, there are currently no programs for the early detection of lung tumors, as there are for other neoplasms (breast, prostate, colorectal, etc.), since computed tomography (CT) scans despite having greater sensitivity and specificity than Simple X-ray for the detection of the pulmonary nodule, administers high doses of radiation, making the risk / benefit in a large population unfavorable. And, given the aforementioned reasons to understand the difficulty of interpreting studies for the characterization of the pulmonary nodule, the simple chest X-ray should be analyzed by specialized personnel in one of the explorations with the greatest interobserver discrepancy, also considerably increasing their workload.

The development of a semi-automatic tool for the detection of pulmonary nodules could be able to detect early stages of the disease, even when the changes in density level in the lung parenchyma are so subtle that human eyes cannot distinguish. To do this, it should be noted that close collaboration is required between radiologists, given that they know the pathophysiological process of the disease as well as its findings in imaging tests, as well as with the developers of the tool, since they have the knowledge to translate the necessary variables and the development and integration of relevant applications.

On a medical level, it would be relevant to detect subtle increases in density in lung parenchyma topographies, in which irregular edges can be discerned and do not correspond to any other anatomical structure. Ideally, once the alteration is detected, it should be located to verify that the radiologist sees the same alteration. For this, a database of negative studies for pulmonary nodules supervised and verified by a radiologist has been collected, to be able to supply a stock of chest X-rays as a sample. After that, face the tool with a group of positive and negative studies for pulmonary nodules and check the success rate, obtaining the results described in the rest of the document.

Once developed and trained, if after analyzing the exploration with the tool, we detect alterations, we could expand the study of the patient with the pertinent tests to stage the neoplasia and establish the opportune early treatment, thus increasing the survival rate. We would thus avoid overlooking alterations that a posteriori condition that the disease is advanced enough, as it often happens, to be able to be treated.

By also integrating it into the image viewer, the tool would act as a second review in all chest X-rays that were carried out, not only in lung cancer screening. The first practitioner would analyze the image as well as the tool. In the event that the tool detects alterations, it would issue an alert on the screen for the physician to request that a radiologist, a third reviewer, analyze the image, issue the corresponding report, and request the tests deemed appropriate if indicated.

With all this, associated with the fact that the radiation applied in a chest X-ray is much lower than that of a tomography, a system could be established for the early detection of lung cancer by screening with annual chest X-rays in patients with risk factors. The benefit would be the early detection of

the disease with its consequences on the prognosis, avoiding overlooking positive cases, as well as providing help for the detection of cases to all doctors by incorporating the tool in the viewer.

But once the tool has been developed and implemented, training for the detection of other pathologies that follow a fixed pattern of representation can continue. Pneumothorax, pleural effusions, signs of heart failure or rupture of thoracic aneurysms, among many other pathologies, present within their variability, relatively frequent and even staged patterns. Therefore, by integrating the tool into the information system of a center, and the verification of studies by a radiologist assigning the findings to be detected of said pathology, the number of pathologies to be detected by the tool could be expanded. In addition, it is continuously improving by nourishing itself with all the explorations that are carried out on a daily basis, progressively increasing its effectiveness, and further improving the help it can provide to physicians of all specialties in the detection of a wide range of pathologies.

Currently a large number of X-ray imaging studies accompanied by radiological reports are accumulated and stored in the image archiving and communication systems (PACS) of many modern hospitals (Tensorflow 2021, April 4). On the other hand, it is still an open question how this kind of hospital-sized knowledge database containing invaluable image computation (i.e., freely labeled) can be used to facilitate data-hungry deep learning paradigms in building high-precision truly large-scale computer-aided diagnostic (CAD) systems.

One of our aims in this paper is to obtain a rich and comprehensive dataset for performing screening tasks (not an exact detection of the disease). Currently, there is no tool that allows a screening of suspicious cases and that refers to them to the supervision of a radiologist, since all cases that may go through an emergency room and the lack of experts in radiodiagnosis lead to many X-rays are not reviewed by a radiologist.

Our research hypothesis is that by selecting a suitable and proper dataset for training a convolutional neural network (CNNs) we can improve the effectiveness in screening tasks as those previously described.

To this end, a joint project has been developed between the UJI (Jaume I University) and the Castellón General Hospital, which has been previously authorized by the Castellón General Hospital CEIM.

The main objectives of the intended dataset are the following:

- Possibility of detecting suspected cases of lung cancer, which a non-imaging physician might miss.
- Train CNNs to detect nodules that a radiologist cannot easily visualize but contain patterns suitable for a CNN.
- Combine with other existing datasets in order to enhance its contrastive power for training CNNs.

2. RELATED WORK

In this section we revise the main approaches in both dataset generation for automatic radiodiagnosis and the state-of-the-art algorithms for lung nodule detection which are all based on deep learning.

2.1. CNN for Computer Vision

CNN networks currently cover the state-of-the-art methods for image classification. CNN consists of a series of layers that perform two basic operations over the images: convolution and max-pooling. Convolutions are applied by means of a series of learnable filters of a given size (i.e., kernels) which automatically extracts features from the images. Max-pooling layers are aimed at summarizing the results of the convolution layers by applying another kernel with the max-pool layer. For image classification, the number of layers are usually large, between 50 and 112, which implies a high number of parameters to train. Moreover, a final fully connected layer is put at the end of the CNN in order to classify the images into the intended classes.

In the literature several configurations of CNNs have been proposed for image classification, like ResNet, AlexNet, and many others. Basically they defined the main hyperparameters of the CNN: number of layers, size of kernels, and so on.

Due to their large number of parameters to be learnt, the process of training a CNN needs a large number of samples per class. The usual way to alleviate this problem is to do transfer learning. With this method a CNN trained with a huge amount of images is re-trained with a small number of samples in the target domain (e.g., X-ray images). Transfer learning allows us to reuse many of the image patterns captured in extensively trained CNNs for classifying images specific to the target domain.

In the field of radiodiagnosis, the images come in DICOM format and are usually of high resolution. Current CNN architectures support a maximum of 1024x1024 in Tensorflow (Tensorflow 2021, April 4) and 4096x4096 in Pytorch (Nvidia 2021, April 4). The experiments that are explained later have been carried out from 224x224 to 512x512 resolutions, the latter having been the best results with a combination of a growth rate of adaptation to the images. Which indicates that the resolution in medical fields matters.

Convolutional neural networks, hereinafter CNN, were introduced in the 90s of the last century (LeCun, Y. et al., 1990, 396–404). The CNN allows to reveal a connection structure between data through the journey of different layers and these are composed of neurons that will react to the data through an algorithm, which will obtain a classification of the data with respect to the responses of the neurons and it will reveal connections that otherwise could not have been discovered (LeCun et al., 2015).

Some recent improvements to basic CNN include the following approaches (Karim, 2019): combining different kernel sizes in the same convolution (start) layer, deepening them, and feedback (Karim, 2019) (Ananthaswamy, 2021). An attempt is made at all times to understand the functioning of the brain in order to use it and improve CNN (Ananthaswamy, 2021).

2.2. Datasets for Automatic Radiodiagnosis

In the literature we can find some contributions aimed at providing datasets from existing RIS/PACS systems. ChestX-ray8 (Wang et al., 2017) provides 108,948 frontal view X-ray images of 32,717 unique patients with the image labels of eight diseases extracted from associated radiological reports by applying natural language processing (NLP). In this case, each image can have multiple tags. Importantly, this work demonstrates that common thoracic diseases can be effectively detected, and even spatially located, through a unified multi-label image classification and disease location framework, even though these have been poorly monitored. This will be one of the main starting hypotheses of this work. However, while the initial quantitative results of that work are promising, the proposed CNN relies on "reading chest radiographs", recognizing and locating common disease patterns but still far from a fully automated high-precision CAD system.

Authors of (Gozes & Greenspan, 2019) also propose CheXNet, an algorithm that can detect pneumonia from chest radiographs at a higher level than practicing radiologists. The algorithm consists of a 121-layer CNN. This network was trained on ChestX-ray14, currently the largest publicly available chest radiography dataset, which contains more than 100,000 frontal-view X-ray images with 14 diseases. Four practicing academic radiologists write down a set of tests, comparing CheXNet's performance with that of radiologists. Results showed that CheXNet exceeds the average radiologist performance on the F1 metric. Finally, they extended CheXNet to detect all 14 diseases in ChestX-ray14, achieving cutting-edge results across all 14 diseases.

Another interesting recent work related to provision of datasets in the field is PADCHEST (Bustos et al., 2020). PADCHEST also gets the image labels by extracting concepts from reports with a recurrent neural network (RNN). This dataset includes more than 160,000 images obtained from 67,000 patients that were interpreted and reported by radiologists at Hospital San Juan Hospital (Spain) from 2009 to 2017, covering six views from different positions and additional information on image acquisition and demographics of the patient. The reports were labeled with 174 different radiographic

Table 1. Summary of proposed datasets related to chest X-ray diagnosis

Dataset	Size	Classes	Normality Class
NCCIH	112,120 images	14 thoracic diseases (nodules ~ 6,323)	Yes
OpenI	7,470 images	Thoracic diseases (nodules ~ 106)	?
CheXPert	224,316 images	Multiple annotations (no nodules)	No
PADChest	166,000 images	Multiple annotations (nodules ~ 6,084)*	Yes
PLCO Lung Dataset	185,421 images	Lung cancer	-
MIMIC-CXR	377,110 images	Multiple annotations (no nodules)	-
JSRT	247 images	malignant-benign	Yes (38%)
Mendeley	1824 images	Covid-19 (no nodules)	Yes (912)

findings, 19 differential diagnoses, and 104 anatomical locations organized as a hierarchical taxonomy and mapped in the standard terminology of the Unified Medical Language System (UMLS). 27% of the reports were manually annotated by trained physicians, and the remaining set was labeled using a supervised method based on the proposed RNN. To our knowledge, this is one of the largest public chest X-ray databases, suitable for training supervised models on radiographs, and the first to contain radiographic reports in Spanish.

Despite claims that they achieve and/or exceed performance at the medical level, current deep learning models for the classification of pathologies using chest X-rays are showing that they are not generalizable in all institutions and are not ready yet for adoption in real-world clinical settings (Zech et al., 2018). Furthermore, warnings of possible unintended consequences of its use are discussed in (Cabitza et al., 2017).

All the datasets mentioned above also suffer from a critical mass to be able to train a neural network in a convenient way as in the case of ImageNet (Imagenet 2021, April 4). This is due to the fact that correctly labeling an X-ray image involves highly qualified personnel, not as in ImageNet, where any volunteer can label images.

2.3. CNN Models for Radiodiagnosis

Main CNN approaches for radiodiagnosis rely on the previous research of CNN with ImageNet. Indeed, many approaches fine tune pre-trained ImageNet networks to medical imaging (transfer learning) in order to deal with the small labelled datasets available so far. As mentioned earlier, the recent proposal of large datasets of X-ray labelled images has allowed researchers to better adapt CNN architectures to radiodiagnosis. In (Olsen et al., 1996) authors propose a specific CNN called DenseNet-121 specially designed for the ChestXray14 dataset. This dataset included 14 common thoracic diseases, including lung nodules. Then, by applying transfer learning, specific models for some diseases can be defined. Authors reported good results when fine tuning DenseNet-121 from small datasets for tuberculosis.

Computer-aided detection (CADe) (Park et al., 2019) has greatly benefited from recent advances in convolutional neural networks. These networks have shown to be very useful in detecting tumors either benign or malignant, reducing the need of invasive procedures and preventing errors in diagnosis. However, designing an efficient network configuration for a particular classification problem is an arduous and arcane task. Some work like (Xiaoxuan et al., 2019) is focused on searching optimal

Table 2. CNN models for nodule detection

Approaches	Dataset	Network Configuration	Metrics
ChestX-ray8 (Wang et al., 2017)	NCCIH	Framework with 4 pre-trained CNN models (ResNet-50, VGG, AlexNet and GoogLeNet)	nodules 0,71 AUC
ChestX-ray14 (Xia et al., 2020)	NCCIH	Same as ChestX-ray8	nodules 0,67 AUC
CheXNet (Gozes & Greenspan, 2019)	NCCIH	CNN with 121 layers trained with ChestX-ray14 data	nodules 0,78 AUC pneumo 0,43 AUC
Genetic DL (Genome 2021, April 4)	PLCO	DeepNEAT-Dx (optimal CNN)	97.15% accuracy
Optimized and improved Faster R-CNN (Bhandary et al., 2020)	LIDC-IDRI	Fast R-CNN	91.2% accuracy
DENSENET (Xuechen et al., 2020)	JSRT	Densenet	99% accuracy

CNN architectures. In this work, authors propose a genetic algorithm to find an optimal CNN aimed at finding early-stage lung cancer on radiographs (CXR), which achieves 97% of accuracy with the best found configuration. For the same dataset, RestNet-152 achieved 92% of accuracy. This result indicates that the selected dataset is not too hard compared to larger collections where detecting nodules is a hard task. Something similar occurs with other recent work like (Bhandary et al., 2020) and (Xuechen et al., 2020).

Another issue that has been scarcely treated in the literature is the visualization or explanation of the predictions of a CNN. For example, authors in (Liu 2019, October 22) propose a method to detect and localize the parts of the radiographs related to pulmonary tuberculosis.

3. MATERIAL AND METHODS

For the screening experiments we mainly use two large datasets of chest X-ray images associated where lung nodules are present. More specifically, we use Chest-X-ray8 (Xiaosong et al., 2017) and PadChest (Bustos et al., 2020). Both datasets were described in Section 2.2. Apart from these datasets we used other datasets related to X-ray for extracting samples for the normality class.

Our ultimate goal is the creation of a large and balanced lung cancer dataset that contains X-ray images with nodules difficult to detect by a non-expert physician. Images should have associated reports where lung cancer is diagnosed with CT. In addition, X-ray images are requested before the disease is declared, since they will be used to look for patterns in images where the findings are too subtle to be detected by the human eye. For this purpose we added X-ray images from the Hospital of Castellón fulfilling these constraints.

3.1. Validation and Precision

To avoid what has happened in the study "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis" (Xiaoxuan et al., 2019) in the Lancet of September 25, 2019 that is cited "Our review found that the diagnostic performance of deep learning models is equivalent to that of healthcare professionals. However, an important finding of the review is that few studies presented externally validated results or compared the performance of deep learning models and healthcare professionals using the same sample. Additionally, poor reporting is common in deep learning studies, limiting reliable interpretation of the reported diagnostic accuracy. New reporting standards that address specific deep

learning challenges could improve future studies, allowing greater confidence in the results of future evaluations of this promising technology. "

To do this, the corresponding calculations of the rates and errors will be made, which will be explained below. But, the results will be supported by a radiologist who will monitor at all times if the results are correct.

For the calculation of the success and error rates, the classification problem (Olsen et al., 1996) is one of the first to appear in scientific activity and constitutes a process inherent to almost any human activity, in such a way that in problem solving and in decision-making, the first part of the task consists precisely in classifying the problem or situation, and then applying the corresponding methodology, which will largely depend on that classification. Of course, this is also the case in medicine, a science in which diagnosis is an essential part, being a preliminary phase for the application of therapy. Diagnosing is equivalent to classifying a subject in a specific pathology based on the corresponding data from her anamnesis, examination and complementary tests. When we talk about classifying a subject in a given group, based on the values of a series of parameters measured or observed, and that classification has a certain degree of uncertainty, it is reasonable to think of the use of a probabilistic methodology, which allows us to quantify that uncertainty.

This uncertainty, to several types of errors (Dekking, 1946): the type I error, also called type alpha error (α) or false positive, is the error that is committed when the researcher rejects the null hypothesis (H0: the initial assumption) being this true in the population (in statistics a population is a set of objects, individuals, elements or events with certain characteristics). It is equivalent to finding a false positive result, because the researcher concludes that there is a difference between the hypotheses when in fact there is not. It is related to the level of statistical significance.

The hypothesis from which H_0 is started here is the assumption that the experimental situation would present a "normal state." If this "normal state" is not observed, although it actually exists, it is a statistical type I error. The example for type I error would be:

The patient is considered to be sick, even though he is actually healthy; null hypothesis: The patient is healthy.

In a research study, type II error, also called beta (β) error (β is the probability that this error exists) or false negative, is committed when the researcher does not reject the null hypothesis, being this false in the population. It is equivalent to the probability of a false negative result, since the researcher concludes that he has been unable to find a difference that exists in reality.

The power or potency of the study represents the probability of observing in the sample a certain difference or effect, if it exists in the population. It is the complement of the type II error (1-β).

A null hypothesis or base H_0, or the alternative hypothesis H_1, and the chosen decision will coincide or not with the one that is actually true. The four cases that are exposed in the following table can be given.

3.2. Experiments and Results

A series of experiments has been established to validate the hypotheses and approximate the value of p (e.g., p-value) is defined as the probability corresponding to the statistic if possible under the null hypothesis. If it meets the condition of being less than the arbitrarily imposed level of significance, then the null hypothesis will eventually be rejected. (value of the calculated statistic). Which in medicine means approaching $p = 0.05$.

As previously mentioned, to measure screening results, we apply ROC curve (Wikipedia 2021, April 4), which illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or probability of detection in machine learning.

Table 3. Error types

	H_0 is true	H_1 is true
It was chosen H_0	No mistake (1-α or true positive)	Type error II (β or false positive)
It was chosen H_1	Type error I (α or false positive)	No mistake (1-β ortrue negative)

The sensitivity, recall, hit rate, or true positive rate (TPR) is defined as follows:

$$TPR = TP / P = TP / TP + FN = 1 - FNR$$

Fall-out or false positive rate (FPR) is defined as follows:

$$FPR = FP / N = FP / FP + TN = 1 - TNR$$

where TP is the number of true positives, TN accounts for the number of true negatives, FP accounts for false positives and FN for false negatives.

3.2.1 Experiments for Pneumonia Detection

For the experiments we used FastAI (https://www.fast.ai/), which relies on PyTorch, a popular library for vision and natural language processing with deep neural networks.

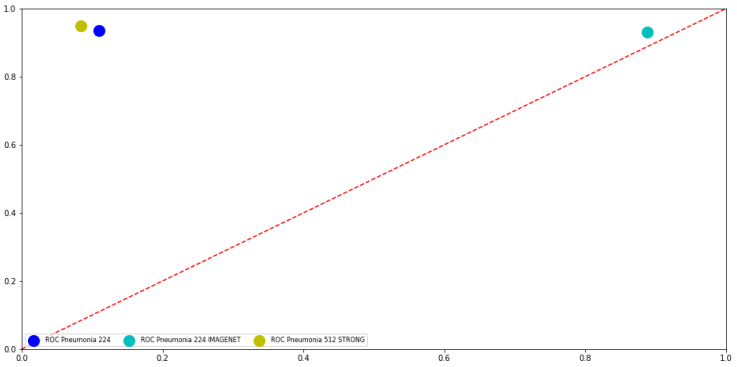
Kaggle's data were used for pneumonia training (Kaggle 2021, April 4), which were selected because the data were on a gray scale as they were chest radiographs with pneumonia. You are, they will appear a lot on X-rays with nodules that will be tried to detect in the next experiment.

Pytorch's Model sub-package (Pytorch Models 2021, April 4) contains model definitions to address different tasks, including: image classification, semantic pixel segmentation, object detection, instance segmentation, person keypoint detection, and video classification. For these reasons, a model was chosen to be used at FastAI with the utmost precision as Resnet-152. Although it is expensive when training, maximum precision is required because a patient is being diagnosed.

When the Resnet-152 model was chosen, it was decided to perform several experiments. In the first one, we used a resolution of 224 pixels and the pre-trained weights from Imagenet. The result was not satisfactory because Imaginet, although it already includes X-rays, they are not exhaustive enough. This experiment corresponds to the green point in Figure 1.

The second experiment was carried out with a resolution of 224 pixels and without pre-trained weights. Then, the expected success was achieved as seen in the graph of the ROC curve of "PNEUMONIA DETECTION" (blue point). The results are even better with a resolution of 512 pixels (yellow point).

Figure 1. ROC curve for pneumonia detection (Y = TPR, X = FPR). The yellow dot corresponds to a resolution of 512 pixels, and the blue one to 224 pixels. The green dot experiment uses a resolution of 224 pixels and pre-trained weights of Imagenet.



3.2.2 Experiments for Nodule Detection

Using the techniques learned in the pneumonia detection, we carried out several experiments for nodule detection. With this purpose, we used the Resnet-152 model trained from scratch with a resolution of 512 pixels.

First experiment: We trained Resnet with only NCCIH data. This model achieved an accuracy of 79% during training, but it did not classify test data (all samples are assigned to the nodule class). Using the pre-trained weights of Imagenet, the accuracy improves up to 82% during training, but test data is all again assigned to the nodule class. This experiment is shown in Figure 2 with a yellow dot.

Second experiment: After analyzing the sample images, we concluded that samples for normality were not good enough for training. Therefore we used the normality samples from the pneumonia dataset. This led to the first success in detecting nodules as well the normal samples from NCCIH. The resulting model behaves correctly on the upper left margin of the ROC curve (purple dot in Figure 2). However, the model exhibited a poor performance when classifying nodule images from other datasets.

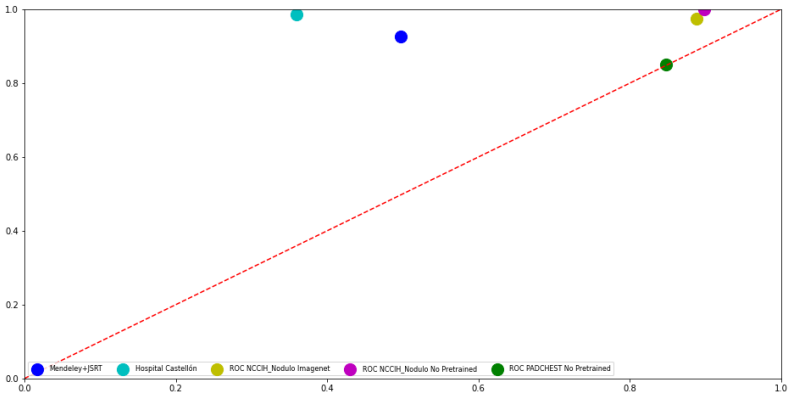
Third experiment: We attempted to generalize the model in order to classify images from different test datasets. With this purpose we sampled data from the network of hospitals of the Generalitat Valenciana (GVA). We added new normality samples to those of NCCIH, and re-trained the model with these new data. Results showed a notable improvement according to the ROC curve (green dot in Figure 2).

Fourth experiment: We introduced more normality samples where the images did not include probes, catheters, electrodes or other pathologies. With these samples, the ability to discern nodules improved. We also added to the training dataset NCCIH samples (of both classes) as well as PADCHEST images with nodules. As a result, the model produced quite good results (cyan dot in Figure 2).

Fifth experiment: To corroborate that the neural network deals with images from other datasets, another experiment was performed over a Mendeley dataset (Mendeley 2021, April 4). In this case, the results were quite good (dark blue dot in Figure 2).

To conclude, these experiments show how the selection of good samples from existing datasets can lead to a good classification model. Resolution of the images is also an important issue, but its impact is minor compared to the selection of samples. We also tested other CNNs like VGG16,

Figure 2. ROC curve for nodule detection (Y = TPR, X = FPR)



Resnet-18 and Resnet-152, but results were similar or even worse. We also did not find any difference in using TensorFlow or Pytorch libraries.

4. CONCLUSION

In this paper we have introduced the problem of screening radiological images for alleviating the workload of the radiology department as well as for the early detection of diseases like lung cancer. We have focused our experiments in the current state-of-the-art methods, all of them based on deep networks of the CNN family. We have revised the main approaches in the literature as well as all the current datasets that are available for training these networks.

Our experiments show that we are still far from an optimal method for automatic screening of X-ray images. However, results demonstrate that a proper selection of samples for the target classes can lead to successful classification models. Moreover, the availability of on-line resources such as those provided by the GVA administration allows us to increase notably the quality and quantity of samples for the target task, and hence the quality of the models.

Our best model for screening allows us to filter out 70% of the X-ray images with a true positive rate near 100%. This implies a good screen ratio and consequently a much lower workload for radiologists. Nevertheless, these results must be further corroborated with a much large-scale experiment involving several hospitals from the GVA network.

As future work, we plan to explore other alternative ways to train networks like Generative Adversarial Networks (GAN), which has shown promising results for medical images. We also plan to integrate image and associated report texts by means of the emerging transformer models like CLIP or ViT in order to get proper attention mechanisms on the image according to the target classes.

ACKNOWLEDGMENT

This research was supported by the “Departamento de Lenguajes y Sistemas Informáticos” of the Universitat Jaume I (UJI), and the UJI research support project with contract number 19I183.

REFERENCES

A Genome Wide Scan of Lung Cancer and Smoking. Dataset. (2021, April 4). https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000093.v2.p2

Abhir Bhandary, G. (2020). Deep-learning framework to detect lung abnormality – A study with chest X-Ray and lung CT scan images. *Pattern Recognition Letters*, *129*(January), 271–278.

AnanthaswamyA. (2021, April 4). https://www.investigacionyciencia.es/noticias/las-redes-neuronales-dan-por-fin-pistas-de-cmo-aprende-el-cerebro-19596

Bustos, A., Pertusa, A., Salinas, J.-M., & de la Iglesia-Vayá, M. (2020). PadChest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis, 66*.

Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *Journal of the American Medical Association*, *318*(6), 517–518.

Dekking, M. (1946). A modern introduction to probability and statistics: Understanding why and how. Springer.

Gozes, O., & Greenspan, H. (2019). Deep Feature Learning from a Hospital-Scale Chest X-ray Dataset with Application to TB Detection on a Small-Scale Dataset. *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),* 4076-4079. doi: doi:10.1109/EMBC.2019.8856729

Imagenet. (2021, April 4). http://www.image-net.org/

Kaggle. (2021, April 4). https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

KarimR. (2019, July 29). https://towardsdatascience.com/illustrated-10-cnn-architectures-95d78ace614d

LeCun, Y. (1990). Handwritten digit recognition with a back-propagation network. Proc. Advances in Neural Information Processing Systems, 396–404.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

Li, X., Shen, L., Xie, X., Huang, S., Xie, Z., Hong, X., & Yu, J. (2020). Multi-resolution convolutional networks for chest X-ray radiograph based lung nodule detection. *Artificial Intelligence in Medicine*, *103*(March), 101744.

Liu, Faes, Kale, Wagner, Fu, & Bruynseels. (2019). *A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis.* https://www.thelancet.com/journals/landig/article/PIIS2589-7500(19)30123-2/fulltext

Liu, J., Liu, J., Liu, Y., Yang, R., Lv, D., Cai, Z., & Cui, J. (2019, October 22). *A Locating Model for Pulmonary Tuberculosis Diagnosis in Radiographs*. https://arxiv.org/abs/1910.09900

Mendeley. (2021, April 4). https://data.mendeley.com/datasets/2fxz4px6d8/4

ModelsP. (2021, April 4). https://pytorch.org/docs/stable/torchvision/models.html

Nvidia. (2021, April 4). https://github.com/NVIDIA/pix2pixHD

Olsen, S. F., Martuzzi, M., & Elliott, P. (1996). Cluster analysis and disease mapping—Why, when, and how? A step by step guide. *BMJ (Clinical Research Ed.)*, *313*, 863. doi:10.1136/bmj.313.7061.863

Park, H., & Monahan, C. (2019). Genetic Deep Learning for Lung Cancer Screening. *Conference on Machine Intelligence on Medical Imaging, 2019*. arXiv:1907.11849

Tensorflow. (2021, April 4). https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/tf2_detection_zoo.md

Zech, Badgeley, Liu, Costa, Titano, & Oermann. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine*, *15*(11), e1002683.

Wang, H., Jia, H., Lu, L., & Xia, Y. (2020, February). Thorax-Net: An Attention Regularized Deep Neural Network for Classification of Thoracic Diseases on Chest Radiography. *IEEE Journal of Biomedical and Health Informatics*, *24*(2), 475–485. doi:10.1109/JBHI.2019.2928369

Wang, X., Peng, Y., Le Lu, Z. L., Bagheri, M., & Summers, R. M. (2017). *ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases*. IEEE CVPR. https://nihcc.app.box.com/v/ChestXray-NIHCC/

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3462-3471. doi: doi:10.1109/CVPR.2017.369

Wikipedia. (2021, April 4). https://en.wikipedia.org/wiki/Precision_and_recall

Rafael Berlanga is full-time professor of Computer Science at Universitat Jaume I, Spain, and the leader of the Temporal Knowledge Bases research group. He received the B.S. degree from Universidad de Valencia in Physics, and the PhD degree in Computer Science in 1996 from the same university. His current research interests include text mining, knowledge bases, information retrieval, and the semantic web. He has led several research projects, has published more than 20 contributions to high-impact international journals and more than 50 contributions to international conferences.