Future Generation Computer Systems 144 (2023) 291–306

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Deep learning for understanding multilabel imbalanced Chest X-ray datasets

Helena Liz [a,b,*], Javier Huertas-Tato [a], Manuel Sánchez-Montañés [c], Javier Del Ser [d,e], David Camacho [a]

[a] Computer Systems Engineering Department, Universidad Politécnica de Madrid, Alan Turing s/n, Madrid, 28031, Spain
[b] Department of Computer Sciences, Universidad Rey Juan Carlos, Tulipán s/n, Móstoles, 28933, Spain
[c] Computer Science Department, Universidad Autónoma de Madrid, Madrid, 28049, Spain
[d] TECNALIA Basque Research & Technology Alliance (BRTA), P. Tecnologico 700, Derio, Bizkaia, 48160, Spain
[e] University of the Basque Country (UPV/EHU), Bilbao, 48013, Spain

ARTICLE INFO

Article history:
Received 29 July 2022
Received in revised form 28 December 2022
Accepted 4 March 2023
Available online 6 March 2023

Keywords:
Convolutional neural networks
Chest X-rays
Explainable AI
Ensemble Methodology

ABSTRACT

Over the last few years, convolutional neural networks (CNNs) have dominated the field of computer vision thanks to their ability to extract features and their outstanding performance in classification problems, for example in the automatic analysis of X-rays. Unfortunately, these neural networks are considered black-box algorithms, i.e. it is impossible to understand how the algorithm has achieved the final result. To apply these algorithms in different fields and test how the methodology works, we need to use eXplainable AI techniques. Most of the work in the medical field focuses on binary or multiclass classification problems. However, in many real-life situations, such as chest X-rays, radiological signs of different diseases can appear at the same time. This gives rise to what is known as "multilabel classification problems". A disadvantage of these tasks is class imbalance, i.e. different labels do not have the same number of samples. The main contribution of this paper is a Deep Learning methodology for imbalanced, multilabel chest X-ray datasets. It establishes a baseline for the currently underutilised PadChest dataset and a new eXplainable AI technique based on heatmaps. This technique also includes probabilities and inter-model matching. The results of our system are promising, especially considering the number of labels used. Furthermore, the heatmaps match the expected areas, i.e. they mark the areas that an expert would use to make a decision.

1. Introduction

In recent years, the field of medicine has faced two relevant problems that hinder patient care: staff workload and subjectivity in the interpretation of tests [1,2]. These problems have no easy solution, which is especially dangerous in medicine because procedural errors can lead to serious health complications. Firstly, overwork in medicine, aggravated in recent times by the global COVID-19 pandemic, can lead to errors and delays in diagnosis and treatment. As mentioned above, there is also subjectivity in the interpretation of some medical tests. The expert analysing these tests, for example X-rays, may arrive at an erroneous diagnosis due to, for example, the existence of signs of different diseases to different degrees [3]. This type of imaging test is one of the most common in various diagnoses due to its low cost, speed of acquisition and the fact that it does not require much preparation [4]. Chest X-rays are useful for detecting a variety of diseases of the chest related to different organs such as the heart, lungs or bones. The features of X-rays make them suitable for analysis with convolutional neural networks (CNN) [5]. The combination of AI algorithms and medical knowledge can improve the performance of medical staff [6] and could also reduce patient waiting times by speeding up the diagnostic process and reducing the workload of doctors.

CNNs have been a breakthrough in computer vision due to their ability to extract features from images. These architectures are composed of different layers. The first has convolutional layers that are inspired by the notion of cells in visual neuroscience. The architectures are based on the visual cortex of animals. The main reason why these architectures have stood out is their great capacity to extract patterns from data, improving the performance of previous systems based on Machine Learning models. This advantage has made them a benchmark in Deep Learning due to their high performance in a wide range of tasks, such as speech recognition, computer vision or text analysis [7].

* Corresponding author at: Computer Systems Engineering Department, Universidad Politécnica de Madrid, Alan Turing s/n, Madrid, 28031, Spain.
E-mail addresses: helena.liz@urjc.es (H. Liz), javier.huertas.tato@upm.es (J. Huertas-Tato), manuel.smontanes@uam.es (M. Sánchez-Montañés), javier.delser@tecnalia.com (J. Del Ser), david.camacho@upm.es (D. Camacho).

https://doi.org/10.1016/j.future.2023.03.005

The properties of chest X-rays make them susceptible to be analysed by this type of algorithms. Some of the main advantages of CNNs over traditional techniques are that it is not necessary to manually extract image features or perform segmentation, and that by being able to learn from large volumes of data they can identify patterns that are difficult for the human eye to detect. Although in this article we focus on classification problems, other problems can be solved, such as X-ray segmentation [8], localisation, regression (such as predicting drug dosage), among others. CNNs are a potential tool for the analysis of chest radiographs. However, most of the work in this field focuses on binary and multiclass classification problems. Actual problems are usually more complex than the above; they tend to be multilabel classification problems, i.e. the different labels are not mutually exclusive, whereas in binary and multiclass classification problems there is only one label per radiograph [9]. To solve multilabel problems, we need to explore new strategies. Adapting algorithms can interpret this kind of problem by transforming them into simpler problems that can be solved by traditional algorithms, i.e., transforming them into binary problems [10]. In the field of chest X-rays we can find samples without labels, healthy patients and samples with radiological signs of several diseases at the same time. On the other hand, there are a large number of different radiological signs in chest X-rays, so if we want to build and validate a system that approximates realistic conditions, we have to use a dataset with a large number of mutually non-exclusive labels. This is the case of the PadChest database [11], which has 174 different radiological signs, substantially increasing the degree of realism and the complexity of the problem.

Many machine learning algorithms, including CNNs, work best when the classes in the dataset are balanced. However, in real life it is common to find datasets where this condition is not met; they are imbalanced datasets, where one or more classes have substantially more examples than the rest. As a consequence, with such datasets, machine learning algorithms learn a bias towards the majority class, even though the minority class is often more relevant. Therefore, it is necessary to apply different methods to improve the recognition rate [12]. There are several options to overcome this difficulty: (a) modify the dataset, reducing the samples from the majority class or increasing the number of samples from the minority class; (b) modify the algorithms to alleviate their bias towards the majority class, e.g. weighted learners [13]. The problem of unbalanced databases is exacerbated in multilabel classification problems, where multiple minority classes may appear, making this challenge more difficult to solve. In medicine, it is widespread because each disease has a different incidence in the population. Heart disorders top the list of the deadliest diseases, followed by chronic obstructive pulmonary disease, which causes more than 6 million deaths a year. In contrast, other diseases such as lung cancer are the sixth leading cause of death with less than 2 million deaths, according to the World Health Organization.[1] As a result, most radiographic datasets are imbalanced; a clear example is PadChest, the dataset used in this article, where the number of samples in each class approximates the incidence published by the World Health Organization.

These algorithms, like many other Deep Learning and Machine Learning methods, are considered "black box" algorithms because end users can only analyse the input and output, but the inference process is opaque, which reduces confidence in these algorithms. To alleviate this problem, explainable AI techniques have been developed, such as saliency maps, which produce heatmaps that highlight the pixels with the greatest influence on the final prediction [14]. This problem is serious in medicine, where errors can be dangerous for patients [15]. For this reason, explainable AI techniques are essential, as they allow users to understand how the system has arrived at the final result and use it to help diagnose [16]. However, the combination of medical knowledge and AI has many advantages, such as helping to reduce medical errors and speeding up diagnostic processes, leading to improved patient care, as doctors would have more time to attend patients.

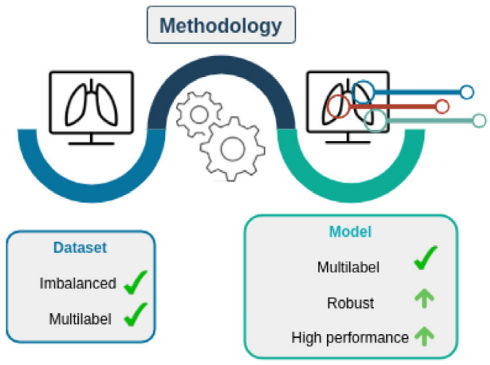

Fig. 1. Visual representation of the problem and the objective of the methodology.

The contribution of this manuscript is a methodology, see Fig. 1, for classifying imbalanced multilabel datasets with many classes The aim of this methodology is to generate robust and quality models; in this case, it has been applied to a highly imbalanced multilabel chest X-ray dataset with 174 classes. We selected this dataset for two reasons: (i) the number of classes, which is higher than in other state-of-the-art datasets; and (ii) the high imbalance between these classes. This methodology will allow to establish a suitable benchmark for this dataset against which future works can be compared, as there are currently very few published contributions using this dataset and they do not provide a detailed analysis of the problem.

We can summarise the main contributions of this work as follows:

- A methodology for imbalanced multilabel classification problems.
- A discussion about the experimental results obtained using a dataset with a large number of classes (more than 30) and a severe imbalance between them.
- An explainability interface using Grad-CAM for multilabel datasets.
- A suitable benchmark for this dataset serving as a reference against which to compare future proposals from the scientific community.

Finally, this manuscript is organised as follows. Section 2 summarises the most relevant work in the literature, with a special focus on chest X-ray classification problems for imbalanced multilabel datasets; Section 3 describes the methodology proposed for this type of problem, consisting of training a model and generating a visualisation based on heatmaps; Section 4 presents the CNN architectures, the hyperparameters used for training, details of the execution environment, and a link to the repository where the code used in the experimentation can be found; Section 5 presents the experimental results, and Section 6 presents the main conclusions and possible lines of future work.

2. Related work

Since the first application of AI techniques in medicine in the 1980s, the use of these algorithms has grown exponentially,

[1] https://www.who.int/es/news-room/fact-sheets/detail/the-top-10-causes-of-death

especially in recent years. Deep learning algorithms are applied to all kinds of clinical data [17]: biosignals, which include electrical [18,19], mechanical [20,21] and thermal signals [22,23]; biomedicine, which studies molecules of biological processes [24–26]; electronic health records (EHR), focused on optimising diagnosis [27–30]; and clinical imaging, widely used in the diagnosis of many diseases [31–34], as is the case with our problem. The practice of healthcare has evolved from observation-based medicine to evidence-based medicine. This makes deep learning and big data algorithms especially useful in this field as they can identify some radiological signs that medical staff cannot detect [35]. Although in this manuscript we focus on classification problems, there are papers where these algorithms are used in regression problems, such as estimating the dose of a drug [36]; generating medical reports from clinical tests [37]; support healthcare management [38]; or image processing, such as image segmentation [39] and image reconstruction [40].

The COVID-19 pandemic has had a strong impact on research into the application of machine learning and deep learning in medical image analysis. As expected, many of the classification systems investigated have focused on detecting signs of bilateral COVID-19-associated pneumonia. In Ahmed et al. [4] they use two different pre-trained architectures to classify chest X-rays, VGG16 and ResNet, and optimise the hyperparameters. In Pham [41] they train three different pre-trained architectures, AlexNet, GoogleNet and SqueezNet, with six datasets independently, testing different percentages of train set samples (50 and 80%), achieving an accuracy of 99.85% with SqueezeNet. However Ahmad et al. [42] develops an ensemble system based on MobileNet and InceptionV3 that achieves 96.49% accuracy. Soon, binary classification was extended to multiclass problems, making it possible to discern whether pneumonia is caused by COVID-19 or another virus/bacteria or whether the patient is healthy. As with binary classification problems, many works, such as Avola et al. [43], use state-of-the-art architectures to find the best performing ones, such as AlexNet, GoogleNet, ResNet and ShuffleNet, among others. MobilNet_v3 achieves the best result with a precision of 84.92% on a dataset composed of 6330 samples. In Zebin and Rezvy [44], in addition to training a pre-trained state-of-the-art architecture, a heatmap-based visualisation is generated that shows two images for each sample. The first is the original X-ray, and the second is the class activation map, i.e. the most important area for the CNN. However, the images do not overlap, making interpretation difficult. Other works, such as Teixeira et al. [45], apply segmentation techniques to remove all irrelevant areas of the system, which should improve performance and visualisation. Their dataset consists of three different classes: COVID-19, normal and lung opacity.

As we have discussed in Section 1, the explainability of deep learning models is a fundamental factor to be taken into account in their application. These models are black-box algorithms and need explainable AI techniques to make them more trustworthy [46]. There are two main ways to produce the final visualisation, (1) generate a heatmap per label, or (2) generate a single visualisation for all classes. The first one is more commonly used, [43,45,47] however, this technique has one main limitation: it is not feasible for a large number of labels, and it makes a global view difficult. The second one (e.g. Teixeira et al. [48]), shows the different signs as areas with higher colour intensity, but only one colour scale was used, which makes it difficult to identify which pathological sign indicates which area of interest. We propose a new technique where each visualisation shows a radiological sign including the probability and agreement between models.

Although most medical datasets have two classes (samples of a particular pathology and healthy samples), in chest X-rays it is common to find signs of more than one pathology. For this reason, in the last five years different authors have published multilabel radiological datasets. These datasets are closer to real situations than binary ones, with the additional challenge of imbalance of different classes. The size of each class in a realistic dataset should depend on the incidence of pathology in society, i.e. some classes are more represented than others. These characteristics of these datasets are interesting and need to be analysed in detail in order to address the problem adequately.

Table 1
Summary table of multilabel datasets in the field of chest radiography.

	# samples	# patients	Labels	Views	Reference
ChestX-ray 14	112120	32717	14	frontal	[54]
CheXpert	224316	65240	14	frontal/lateral	[55]
PadChest	160000	67000	174	frontal/lateral	[11]

2.1. Multilabel classification problems

As we have seen, much of the work generated in recent years has focused on binary and multiclass classification problems. In these problems, the labels are mutually exclusive, while multilabel classification problems have multiple classes that are not mutually exclusive, which increases the difficulty of the problem. There are two ways to solve these problems: (a) transform the multilabel problem into simple binary problems, or (b) adapt the algorithms to solve the multilabel problem directly, i.e. attack the problem globally [49].

Binary and multiclass classification systems are very restrictive, as they only serve to detect one type of radiological finding. However, patients can often present signs of multiple diseases at the same time. There are very few multilabel datasets that take into account a large number of signs, as they require a large number of samples and most of them have only a few labels. Three datasets are worth highlighting for their quality and relevance to the state of the art (see Table 1). The first two have been used extensively in image classification problems, but the third has been used mainly in medical report generation [37,50–53]. However, this third dataset has two advantages that make it very suitable also for classification problems: (1) the number of labels is larger; (2) it has the largest number of different patients, which implies a smaller number of similar samples from the same patient. Given the lack of application of algorithms for this task to increase the potential and interest of this dataset mentioned above, it was selected as a case study for this article.

ChestX-ray 14 dataset is one of the most widely used datasets in the field of chest X-ray classification since its publication in 2019 [54]. For example, Wang et al. [56] uses DenseNet-121 optimising its hyperparameters, obtaining an average AUC of 0.82. The AUC achieved for the pneumonia class was 0.662 (the lowest), while for the hernia class it was 0.923 (the highest). Other researchers use different architectures such as Inception-ResNet_v2 and ResNet152_v2 to achieve an AUC for pneumonia of 0.73 [57]. Much of the work on this dataset retrains state-of-the-art architectures, but there are other strategies for improving classification performance; for example, Almezhghwi et al. [58] switches the classifier from AlexNet and VGG16 to SVM with the intention of improving the results of previous manuscripts, achieving an AUC for pneumonia of 0.98 with both architectures. Having different types of radiographs of the patient can also improve the classification results, for example a frontal and a lateral X-ray. Finally, the main disadvantage of ChestX-ray 14 dataset is that it only contains radiographs with a frontal view, while CheXpert and PadChest datasets also contain X-rays with a lateral view.

The second dataset, **CheXpert** has 14 different labels and reports on all images. In terms of published classification work, we find a situation similar to ChestX-ray 14, with many works retraining state-of-the-art architectures, such as Seyyed-Kalantari et al. [59], where they adjust the hyperparameters of DenseNet-121 to optimise its performance. Other authors look for different strategies, such as Cohen et al. [60], where they make two modifications to DenseNet to improve its performance. First, they modify the loss function by assigning weights to the different labels, alleviating the imbalance problem. Second, they modify the threshold for discerning between the presence or not of each label, i.e. the probability at which the class is considered present. However, the CheXpert dataset has the same limitation as ChestX-ray 14: they contain only 14 possible diseases, which represents only a small subset of all possible diseases that may be present in the chest.

Finally, **PadChest** is the most interesting dataset of the three in our opinion because it has many more labels than the others. It is a massive multilabel classification problem, much closer to reality than the other datasets. The number of patients used is also larger than in the others, leading to more variability in the dataset, and the imbalance of the classes is larger too. One of the papers using this dataset, [61], combines the PA and lateral views to predict labels in four different ways: (a) the lateral view is stacked in the second channel of PA X-ray; (b) both views are processed by two CNNs and the combination of them is processed by a fully connected layer; (c) the model input is processed through two separate CNNs, the output is concatenated and passed by two dense layers with an average pooling layer between them; (d) a modification of (c) where two dense layers are added. A major limitation of that paper is that it shows overall results without performing a detailed analysis per label, which prevents comparison with other works in the area. On the other hand, in Pooch et al. [62] CheXNet is retrained, which is a state-of-the-art architecture previously trained with a multilabel chest X-ray dataset. In that paper, different models are trained with four datasets, and each model is tested with each dataset separately. The main limitation of that manuscript is the reorganisation of the labels of PadChest dataset: the label "Lesion" is generated to unify the samples of the atelectasis classes, using only 8 classes out of the 174 available. Given the limitations we have found in all classification works using the PadChest dataset and that some most of them are not replicable, we propose to create a benchmark that future works can use to compare results, with a methodology adapted for the two main problems: the high number of different labels, and the imbalance between them. As we have explained in this section, the PadChest dataset has several advantages over other multilabel datasets: (i) it has the most labels, which makes it closer to real-world scenarios; (ii) the number and diversity of patients is greater; and (iii) it contains lateral and frontal radiographs. We propose two ways of organising the dataset based on the term tree provided by its authors, which allows us to group radiological signs into higher classes. The first one uses the specific labels for a finer-grained classification. The second one works with more general labels, which indicate more general radiological signs.

2.2. Class imbalance in deep learning

As explained above, most machine learning algorithms work best when the number of samples for each class is similar. When there is a significant difference between the classes, the system will boost the majority class while the minority class(es) will have less relevance, even though the minority class is often the most relevant. There are several classification tasks with this problem, such as Cohen et al. [63], where the majority class is COVID-19 over the rest of the pneumonia classes. As expected, because the incidence of COVID-19 has been extremely high, the dataset contains more than 400 samples of COVID-19 followed by the class *Pneumocystis spp* with fewer than 30 samples. This phenomenon appears in many classification tasks, especially those with more than two classes, both multiclass and multilabel. For example, in Wang et al. [54] there are 15 classes and the class "No findings"/"Normal" exceeds 50,000 samples, while the other labels have less than 20,000 samples, of which only three exceed 10,000 samples.

As mentioned in the Introduction section, there are different strategies to alleviate the class imbalance problem. *Modify the dataset*, for example with oversampling techniques, which increase the number of samples from minority classes by applying data augmentation and histogram equalisation techniques [64]. Charte et al. [65] develops a new algorithm, Multilabel Synthetic Instance Generation, for multilabel problems. For each sample, a nearest-neighbour search is performed, the features are extrapolated and the label is generated from them. Another option for generating synthetic samples is to use generative adversarial networks (GANs), i.e. to use deep learning models to produce new samples from the original dataset. Salehinejad et al. [66] uses this method to generate new chest X-rays to balance the different classes. Another strategy for balancing the classes in the dataset is to reduce the samples of the majority of classes. This technique is called undersampling. Typically, random samples are removed from the majority classes, as in Qu et al. [67], where the maximum number of samples in each class is set to balance it. Undersampling is not as widespread as oversampling because Deep Learning systems need a large number of samples, so undersampling may not work.

Another strategy to alleviate class imbalance is to *modify the way the model learns* by increasing the weight of minority classes in learning, thus preventing the model from giving more importance to majority classes. One option is to apply class weights in the loss function that increase the relevance of the minority classes. One example is Rajpurkar et al. [68], which uses the chest X-ray14 dataset to classify the presence or absence of pneumonia. Another example is Monowar et al. [69], where the weighted binary cross-entropy loss function is applied. Ge et al. [70] developed a novel error function, Multilabel Softmax Loss, this method considers the relationship of multiple labels explicitly, the author computes the derivative of the error with respect to each class using the chain rule. In addition they applied it to a system composed of two CNNs combined by a bilinear pooling layer. Teixeira et al. [48] proposes a dual lesion attention network composed of two models, DenseNet-169 and ResNet-152, as feature extractors, after an attention module and average max pooling. The outputs are combined to generate three classifiers. Finally, all classifiers are merged to obtain the final prediction. In addition, they used a variant of the weighted binary cross-entropy loss. To tackle the class imbalance, we propose using weighted cross-entropy with logits using class weights.

2.3. The challenge of imbalance in multilabel classification problems

As we have explained, many real classification problems have two properties that make them difficult to solve: multilabeling and imbalance. Each of these two properties alone makes classification difficult, so together they can be very challenging. In medicine, multilabel and imbalance problems are common because medical staff can find different radiological signs on a chest X-ray, and different diseases do not have the same incidence in the population. All the datasets mentioned in Section 2.1 have both features; however, ChestX-ray 14 and CheXpert have a low number of classes, 14 labels, compared to PadChest [11], which is composed of 174 different labels with a large imbalance: the label

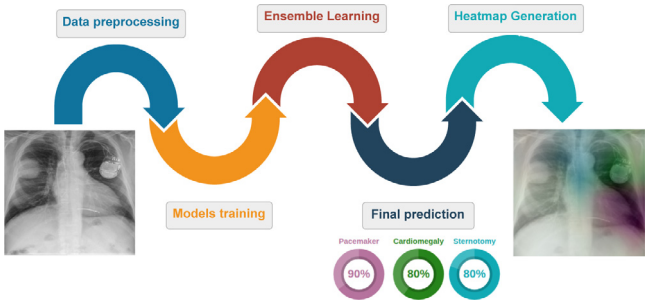

Fig. 2. Visual representation of the proposed ensemble system. We train each architecture with preprocessed images, and their outputs are combined to generate the ensemble output. Finally, the system produces the global prediction and heatmap visualisation.

"Normal" has more than 35000 samples, while other labels, such as round atelectasis, pleural mass or nephrostomy tube, have less than ten samples.

Most of the published work using these datasets modifies the architecture so that it can directly solve multilabel problems, but does not consider or apply any specific technique to solve the imbalance problem. However, other works explore different ways to overcome these difficulties and achieve better results. Such as Huang and Fu [71], which proposes a multi-attention convolutional neural network to reduce the performance difference between classes and, more interestingly, to extract discriminative features to classify similar classes, which is very common in this kind of dataset. Wang et al. [72] generates three images: the first one is the original chest X-ray, the second one is a segmentation-based cropping, where areas not interesting for the model are removed, and the last one is a cropping of the area where previous models have found pathological signs. The information extracted from the three images is fused and finally processed to obtain the final result. Another interesting strategy is the modification of the loss function to focus on the most interesting samples; for example, Qin et al. [73] proposes a loss function called "weight focal loss", which forces the model to pay more attention to the most difficult samples. This makes the model pay more attention to minority classes, avoiding false negatives.

These methods can help in class imbalance problems, but in extreme cases of multilabel and imbalance, such as the PadChest dataset, they may not be sufficient. Most of the published papers attempt to improve the performance of the architecture or solve these problems using a single strategy, which may not be sufficient for datasets such as PadChest.

In contrast to other works in the related literature, we have decided to address these problems by combining different strategies: (1) to avoid confounding the model with areas that do not present interesting radiological signs, we have applied segmentation-based cropping; (2) to make the system robust against the individual errors of the different architectures, we have created an ensemble whose hyperparameters have been adjusted in a validation split to obtain the best possible results; (3) we have applied a specific loss function for imbalanced data that weights each class by its inverse frequency. The combination of these techniques will allow us to substantially reduce the errors due to imbalance and the high number of labels. In addition, we have created a heatmap-based visualisation that highlights the most important areas for detecting each disease represented in the dataset, the estimated probability of that pathology, and the agreement between models (how many models have a probability higher than 50% for that disease), which facilitates interpretation and shows the degree of confidence in the result.

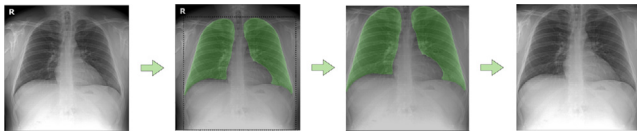

Fig. 3. A segmentation-based cropped sample. The first image corresponds to the original X-ray. The second shows the lung segmentation mask. The third one show the cropped image with lung mask, and finally the last image shows the input of our system, the preprocessing result.

3. Methodology

We can summarise the proposed methodology in Fig. 2, which has four sections. The first is the data pre-processing step, where we prepare the images for the model and apply data augmentation to alleviate class imbalance. In the second stage we build the model, training different state-of-the-art architectures. We then combine the results of each model to obtain the final probabilities. Finally, we developed a multilabel heatmap technique to areas of the image that are relevant in the classification. In this technique, the original X-ray is combined with one or more regions labelled with different colours to facilitate the application of these techniques in health centres or hospitals.

3.1. Label selection

As explained above, multilabel datasets are often imbalanced as they have classes with a low number of samples. For this reason, we must establish a criterion for choosing the labels to include in our classification system, especially in datasets where the number of classes is extremely high, as in our case. First, we set the minimum number of samples a label must have to be included in the classification problem, and we set the threshold at 200 X-rays. For a dataset of 90000 samples this is 0.22% of the total. The model cannot work correctly for under-represented labels as it is not a few-shot system. If a sample has only deleted minority labels we will remove it.

In this paper we consider two different experiments. First, we use the classes proposed by the authors of the dataset that correspond to the specific labels; this classification system has a smaller number of samples and labels due to the cleaning of under-represented labels explained in the previous paragraph, but is a more fine-grained classification system. In the second case, we use more general labels. We create these classes grouping the specific labels according to their characteristics. The number of samples and classes is larger at the cost of being less precise systems, but it allows us to cover a larger number of different classes.

3.2. Preprocessing

The raw images were preprocessed in order to train the model efficiently. First, we reduced the number of channels to one because although the original files are RGB images (three colour channels), the X-rays are grayscale images, so all three channels contain the same information. Next, we normalised their size to 512×512 pixels. The pixel values were then normalised between 0 and 1, Fig. 3 (first image).

Chest X-rays show an area larger than the area of interest (ROI). Areas such as arms or neck, among others, are irrelevant to the problem we want to solve, so a cropping based on segmentation masks was performed, forcing the system to focus on the relevant areas. This trimming is performed in three different steps: first, we generated the lung masks using a segmentation model based on the U-Net architecture [74], Fig. 3 (second image).

We also added the area underneath the lungs to the masks as it may contain radiological signs of interest. On many occasions, the segmentation models are not perfect; they generate more than two masks, leave gaps inside the masks, etc. Therefore, thirdly, we decided to use a mask post-processing system [75]. This system fills the possible gaps in the masks by applying the flood fill algorithm, which analyses the pixels neighbouring the one of interest and depending on whether or not they belong to the mask, it will decide to fill the gap or not. Then, if more than two masks have been generated (one per lung), those whose area is less than a predetermined value are removed. In addition, in case the lung masks are stuck together, they are separated. Finally, the image is cropped using the mask coordinates and the lower boundary of the sample, Fig. 3 (third image). As the images can have different sizes, we normalised their size to 224 × 224 pixels, because this is the normalised size of the samples in the state-of-the-art models, Fig. 3 (last image).

3.3. Image classification with CNNs

Five state-of-the-art architectures pre-trained with ImageNet were selected for their relevance:

EfficientNetB0. [76]: This architecture uses different scaling coefficients to scale width, depth and resolution. In the EfficientNet family, this architecture is the smallest. It is based on the idea that if the images are larger, the network needs more layers to extract the relevant information.

DenseNet-201. [77]: Instead of adding more layers to the architecture, the number of connections between units is increased by connecting each unit to the last, unlike ResNet50, which only connects one unit to the next output. This architecture has several advantages: it alleviates the vanishing gradient problem, enforces feature propagation and feature reuse, and reduces the number of parameters.

InceptionV3. [78]: This architecture is different from the previous ones. It factorises convolutions into smaller convolutions (which can be asymmetric) to reduce cost. In addition, this architecture has an auxiliary classifier between layers that acts as a regulariser.

InceptionResNetV2. [79]: This architecture combines ResNet and InceptionV3. It consists of several Inception units with shortcut connections between them; this enhances the capability of the architecture.

Xception. [80]: It consists of depth-wise separable convolutions involving two steps: depth-wise convolution, which differs from the standard convolution in that it only acts on one channel; and point-wise convolution, where a 1 × 1 convolution is applied to all channels. This architecture also includes shortcut connections, such as ResNet50.

We applied Transfer Learning on the above five architectures and retrained them with PadChest dataset, replacing the classifier in all cases with two dense layers. We froze the first 10% of the convolutional layers, as they detect basic patterns and do not need to be retrained. The remaining convolutional layers are retrained to learn patterns specific to our problem. The main relevant training parameters are summarised in Table 2. In addition, a checkpoint is used to save the best model using the validation loss. Finally, an early stopping algorithm was used to finish training when the validation loss did not improve over the last 25 epochs by more than a threshold of 0.001.

Table 2
Summary of the hyperparameters used in training: optimisation, data augmentation and training methodology.

Optimisation	
Optimiser	Adam
Learning rate	1e−4
Loss	weighted crossentropy with logits
Feed-forward classifier	
# Neurons	512
Activation	ReLu
Dropout	0.2
Data Augmentation	
Shear range	0.1
Zoom range	0.1
Rotation range	45
Width shift range	0.1
Height shift range	0.1
Horizontal flip	True
Fill mode	nearest
Brightness range	0.7–1.1
Channel shift range	0.05
Training methodology	
Maximum epochs	350
Early stopping patience	25
Early stopping threshold	0.001
Batch size	32
Image size	224 × 224

3.4. Ensemble technique

Ensemble learning is an effective way to improve the performance and robustness of deep learning algorithms. We combined the results of all trained models, obtaining a system composed of five different architectures with the same test set. We distinguish two approaches [81]: "Combine then predict" (CTP) and "Predict then combine" (PTC). In the CTP method, the label probabilities predicted by the individual models are first calculated, and then the average probability at each label is used to obtain the ensemble label prediction. The other method, PTC, combines the binary predictions to obtain the ensemble. We consider two versions of PTC: label-wise voting (PTC-lw), which calculates the number of positive and negative individual predictions for each label, adopting the majority. Thus, PTC-lw calculates the prediction of each label independently of the others. On the other hand, PTC-mode calculates the set of labels predicted by each individual model, and predicts the most frequent set.

3.5. Heatmap generation

As explained in Section 2, it is necessary to include XAI techniques for the medical staff to understand the output given by our system. For this reason, we developed a visualisation technique using heatmaps. A heatmap is a matrix of the same size as the input image. The value of each pixel is proportional to its importance for the classification of the model. A colour scale is used in the heatmap to highlight the most relevant pixels for the model.

The first step in generating the heatmaps is to change the activation function of the last layer (the classifier layer) from softmax to linear. Then, for each classifier neuron, we compute the weighted average of the last convolutional layer. Each channel is weighted by the gradient of the classifier neuron with respect to that channel. This is the so-called grad-CAM algorithm [82], which allows to compute a heatmap for each class.

As explained before, the ensemble consists of five models. We generate ensemble heatmaps by averaging the individual heatmaps generated by those models. Finally, we generate a of the average heatmap of each classifier neuron, which is overlaid

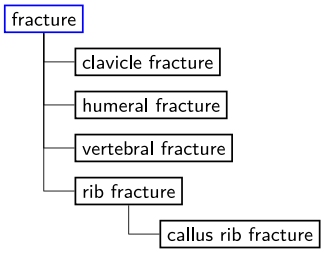

Fig. 4. Example of a section of the term tree of the dataset. The general label is boxed in blue, and the specific labels are marked in black.

Table 3
Summary table of the two types of experiments performed (general labels, and specific labels). The total number of labels and the total number of samples in each of the splits are shown.

	# classes	# samples	Train size	Val. size	Test size
General labels	54	90687	63475	9069	18143
Specific labels	35	85367	59753	8532	17082

on the original X-ray using a 10% of transparency to improve the information for the medical staff. We include in the title the estimated probability for this class and the inter-model agreement showing the confidence of the ensemble in that prediction, which facilitates the use of the system by medical staff.

4. Experimental setup

4.1. Dataset

In this article, we have used the PadChest dataset [11], an imbalanced and multilabel dataset. It was published in January 2019 by the University of Valencia together with BIMCV. The samples were collected at Hospital de San Juan (Spain) between 2009 and 2017. This dataset is composed of 160,868 clinical images from 67,625 patients, divided into 174 different labels, and corresponds to different signs of thoracic disease. This dataset contains chest X-rays with different projections: posteroanterior (PA), anteroposterior (AP) and lateral views; however, only PA X-rays were used for experimentation, corresponding to 91,728 clinical images from the original dataset. The authors of the dataset provided a term tree[2] in which all labels are grouped into more general labels, as can be seen in Fig. 4. In this example, the general label is fracture. The specific labels are clavicle fracture, humeral fracture, vertebral fracture, and rib and callus rib fractures. Therefore, we designed two experiments, the first using specific labels for classification and the second using more general labels, each grouping one or more specific labels. We then set the minimum number of samples that each class must have to be included in the classification system. The more general classification system has a larger number of classes that are more heterogeneous, while the more specific classification system has a smaller number of classes, but is more precise than the previous one.

Table 3 shows the details of the two classification systems, the number of samples, the classes and the size of the training, validation and test sets. In the train/test/validation split we stratify the samples according to classes and patient id, which avoids biases and problems between subsets. In addition, to facilitate the replicability and transparency of this article we will make the split available on the github in Section 4.2.

2 https://github.com/auriml/Rx-thorax-automatic-captioning

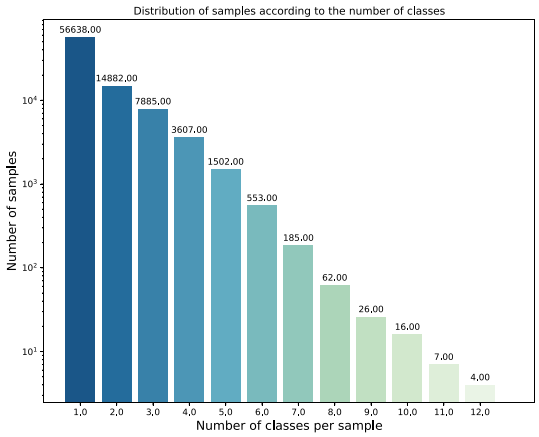

Fig. 5. Distribution of the number of labels per sample (specific labels experiment).

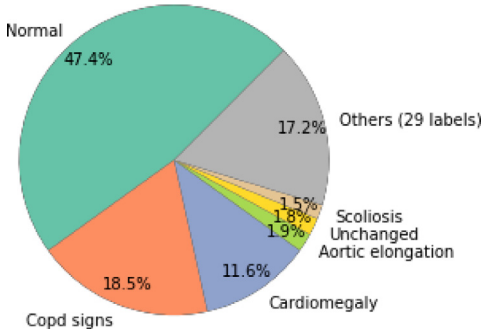

Fig. 6. Label distribution (specific labels experiment).

Label distribution: Specific labels. This experiment, as explained in Section 3.1, label selection, has a smaller number of samples and classes than the second case, but the radiological signs are more accurate. In this experiment we used a total of 85367 samples and 35 different classes. We can observe in Fig. 5, how more than half of the samples present a single class; however, we can observe that there are samples with a high number of classes, four of them presenting 12 different labels at the same time. This distribution of the samples is in line with expectations; the number of samples decreases as the number of labels per sample increases. In Fig. 6 we can see how the classes in this experiment are extremely imbalanced. Although there are 35 classes, the six majority classes account for 82.7% of the dataset. Only the normal class, which is the majority class, accounts for 47.4% of the total samples, while the supra aortic elongation class, which is the least represented class, accounts for only 0.28% of the total.

Label distribution: General labels. In this experiment, different classes were unified according to the tree of terms proposed by the authors. Therefore, the number of classes and samples is higher than in the first experiment. However, the radiological signs used in the classification are less precise, so in the end 54 classes and 90,687 samples were used. In Fig. 7 we can see how the number of classes per sample is distributed in a very similar way to the previous case. However, we can see that there are samples with 13 different labels, one more than in the previous case. If we look at Fig. 8, we can see that the six majority classes represent 51.3% of the total, while the other 48 classes do not reach 50%. The majority class, as in the previous case, is the normal class. This class accounts for 22.6% of the total while the minority class, vascular redistribution, accounts for only 0.13% of

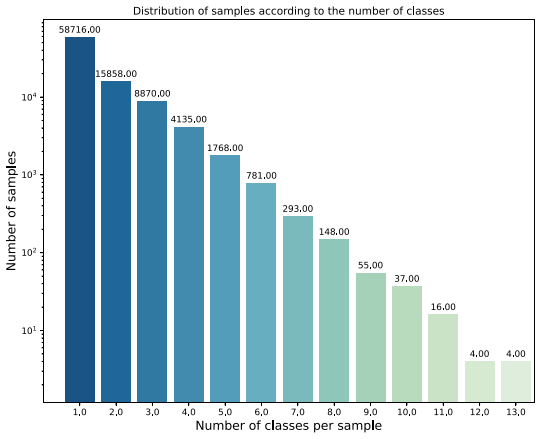

Fig. 7. Distribution of the number of labels per sample (general labels experiment).

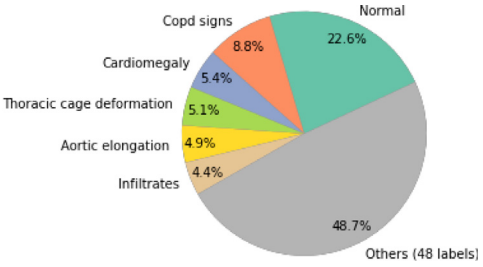

Fig. 8. Label distribution (general labels experiment).

the dataset. This shows that even if we group the radiological signs into higher classes, the dataset is very imbalanced.

4.2. Execution environment and Github repository

All experiments have been run on a 24 GB Nvidia GeForce RTX 3090. The main packages used in these experiments are the following: Tensorflow [83], Scikit-Learn [84] and openCV [85]. The code developed in our work is publicly available at GitHub.[3]

5. Experimental results

This section describes the results obtained with the proposed methodology and evaluates its performance on a multilabel and imbalanced problem, the PadChest dataset. We considered two strategies for the classes: directly using the labels proposed by the dataset creators, or grouping them into more generic classes that encompass similar radiological signs. First, we checked whether preprocessing improves the ensemble performance. Next, we checked the performance of both the individual models and the ensemble, and analyse the quality of the visualisations based on explainable AI techniques. To measure the performance of the different models, we have used three metrics suitable for multi-label problems: Area Under the Curve (AUC), Hamming Loss and F-measure [86].

5.1. Impact of preprocessing techniques

First, we trained the models with the images without segmentation-based cropping or data augmentation. The results obtained, Tables 4 and 5, show that only two individual models have been able to learn, EfficientNet and DenseNet, while the rest of the models were not able to learn and presented a flat training curve with an AUC of 0.5. As expected, the ensemble does not work correctly, and therefore the preprocessing step is necessary.

Tables 6 and 7 show the results training with segmentation-based cropping but without applying data augmentation techniques. At first, it is interesting that Inception does not learn, possibly because it is not able to generalise correctly without data augmentation techniques. InceptionResNet has the best results in most classes, but EfficientNet achieves the best overall result, achieving an AUC of 0.792 while InceptionResNet scores 0.779. Comparing Table 8 with these results shows that the application of data augmentation techniques improves the system performance. If we focus on the results for the different ensembles, we can see that for all labels, the CTP technique performs better than the two PTC methods. CTP also performs better than the individual models except in three cases: in one case it equals them, and in two cases it performs worse). We can conclude that data augmentation improves the performance of the system.

5.2. Performance analysis of CNN models

The first step is the comparison of the different architectures explained in Section 3.3. They are used as a baseline to compare the ensemble system. As explained in Section 3, we consider two types of classification problems: the first uses the original labels proposed by the authors of the dataset ("specific labels"), and the second uses general labels constructed by grouping specific labels. In the first problem, a finer-grained classification is performed, but it contains a small number of labels, 35, as many of the original 144 do not pass the filter of the minimum number of samples (200). In the second problem, general radiological patterns are classified, but there is a larger number of labels, 54, because when grouping labels there are a larger number of classes satisfying the minimum threshold of 200 samples.

Tables 8 and 9 show the results obtained by applying the proposed methodology for the first case study (classification using specific labels). The model with the best global AUC value is DenseNet, followed by EfficientNet, with 0.818 and 0.804 respectively. The other models (Inception, InceptionResNet and Xception) do not achieve an AUC = 0.8. These results are broken down by class. First of all, we can observe that the labels with fewer samples do not show worse results on average than the classes with more samples, which means that we have managed to overcome the data imbalance problems of. It can also be seen in the table that some models perform better with majority classes, such as Inception; others achieve the best results for minority classes, such as EfficientNet and Xception. However, DenseNet 201 and InceptionResNet perform well in both cases.

Secondly, we have analysed the results obtained with the ensemble techniques, using the individual models as baselines. Interestingly, only the CTP technique improves the individual models, as is also the case in Table 6. If we focus on this ensemble technique, we can see that there are two classes, Pleural effusion and pacemaker, where the results of the individual models are not improved. These two classes have 658 and 336 samples respectively, i.e. they are not majority classes, so one hypothesis would be that the ensemble performs worse in minority classes. However, the number of labels for which the ensemble does not outperform the individual models is very small compared to the total. Furthermore, the ensemble achieves an AUC above 0.85 for more than 40% of the labels, which is higher than expected. Since we can observe that the ensemble achieves an AUC higher than 0.9 for classes such as hemidiaphragm elevation, hiatal hernia, or sternotomy, all of them with less than 300 samples, we conclude that class imbalance does not affect our system significantly. Considering that the model is trained for 35 different classes, reaching

[3] https://github.com/helenalizlopez/multilabelimbalancedchestxraydataset

Table 4
Specific labels experiment: results obtained by training the models without segmentation-based cropping or data augmentation. For each label, the individual models with the best performance and the ensembles that outperform all individual models are marked in bold. The best ensemble result is marked in italics unless it ties the random classifier.

	# Samples	DenseNet		EfficientNet		Inception		InceptionResNet		Xception		PTC-mode		PTC-lw		CTP	
		AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Normal	34327	**0.589**	0.470	0.500	0.374	0.500	0.374	0.500	0.374	0.500	0.374	0.500	0.374	0.500	0.374	0.589	0.374
Copd signs	13419	0.500	0.457	0.500	0.457	0.500	0.457	0.500	0.457	0.500	0.457	0.500	0.457	0.500	0.457	0.500	0.457
Cardiomegaly	8412	**0.620**	0.551	0.611	0.563	0.500	0.475	0.500	0.475	0.500	0.475	0.500	0.475	0.500	0.475	**0.633**	0.475
Aorticelongation	1399	0.538	0.509	**0.553**	0.526	0.500	0.479	0.500	0.479	0.500	0.479	0.500	0.479	0.500	0.479	**0.558**	0.479
Unchanged	1311	**0.535**	0.483	0.526	0.504	0.500	0.480	0.500	0.480	0.500	0.480	0.500	0.480	0.500	0.480	**0.543**	0.480
Scoliosis	1073	0.500	0.484	**0.550**	0.522	0.500	0.484	0.500	0.484	0.500	0.484	0.500	0.484	0.500	0.484	**0.550**	0.484
Chronic changes	873	**0.581**	0.481	0.578	0.451	0.500	0.487	0.500	0.487	0.500	0.487	0.500	0.487	0.500	0.487	**0.585**	0.487
Costophrenic angle blunting	703	**0.556**	0.525	0.541	0.532	0.500	0.490	0.500	0.490	0.500	0.490	0.500	0.490	0.500	0.490	0.545	0.490
Air trapping	663	0.500	0.490	0.498	0.510	0.500	0.490	0.500	0.490	0.500	0.490	0.500	0.490	0.500	0.490	0.498	0.490
Pleural effusion	658	0.655	0.573	**0.656**	0.567	0.500	0.490	0.500	0.490	0.500	0.490	0.500	0.490	0.500	0.490	**0.676**	0.490
Pneumonia	651	0.626	0.556	**0.629**	0.566	0.500	0.490	0.500	0.490	0.500	0.490	0.500	0.490	0.500	0.490	**0.645**	0.490
Interstitial pattern	594	**0.597**	0.544	0.582	0.547	0.500	0.491	0.500	0.491	0.500	0.491	0.500	0.491	0.500	0.491	0.594	0.491
Infiltrates	591	**0.615**	0.540	0.594	0.542	0.500	0.491	0.500	0.491	0.500	0.491	0.500	0.491	0.500	0.491	0.612	0.491
Laminar atelectasis	578	0.500	0.491	**0.508**	0.491	0.500	0.491	0.500	0.491	0.500	0.491	0.500	0.491	0.500	0.491	0.508	0.491
Vertebral degenerative	575	0.500	0.491	**0.573**	0.485	0.500	0.491	0.500	0.491	0.500	0.491	0.500	0.491	0.500	0.491	**0.573**	0.491
Kyphosis	526	**0.602**	0.558	0.538	0.520	0.500	0.492	0.500	0.492	0.500	0.492	0.500	0.492	0.500	0.492	**0.606**	0.492
Apical pleural thickening	469	0.500	0.493	0.499	0.488	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.499	0.493
Vascular hilar enlargement	463	0.584	0.510	**0.587**	0.475	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	**0.602**	0.493
Fibrotic band	449	0.500	0.493	0.489	0.484	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.489	0.493
Nodule	449	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493
Calcified granuloma	388	0.500	0.494	0.499	0.494	0.500	0.494	0.500	0.494	0.500	0.494	0.500	0.494	0.500	0.494	0.499	0.494
Callus rib fracture	360	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495
Pacemaker	336	0.627	0.543	**0.646**	0.523	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495	***0.663***	0.495
Aortic atheromatosis	318	0.500	0.495	**0.616**	0.457	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495	*0.616*	0.495
Volume loss	294	0.500	0.496	**0.512**	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	*0.512*	0.496
Sternotomy	292	0.530	0.517	**0.539**	0.506	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	***0.545***	0.496
Bronchiectasis	290	0.500	0.496	0.480	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.480	0.496
Hiatal hernia	287	0.500	0.496	**0.533**	0.506	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	***0.533***	0.496
Pseudonodule	275	0.500	0.496	0.498	0.500	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.498	0.496
Hemidiaphragm elevation	254	0.515	0.496	**0.531**	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	***0.544***	0.496
Alveolar pattern	248	**0.664**	0.531	0.663	0.503	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	***0.695***	0.496
Increased density	239	0.528	0.513	**0.536**	0.502	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	***0.547***	0.496
Vertebral anterior compression	214	0.546	0.510	**0.548**	0.487	0.500	0.497	0.500	0.497	0.500	0.497	0.500	0.497	0.500	0.497	***0.559***	0.497
Suture material	210	0.500	0.497	**0.542**	0.509	0.500	0.497	0.500	0.497	0.500	0.497	0.500	0.497	0.500	0.497	*0.542*	0.497
Supra aortic elongation	200	0.500	0.497	**0.503**	0.497	0.500	0.497	0.500	0.497	0.500	0.497	0.500	0.497	0.500	0.497	***0.504***	0.497
Global		0.543	0.508	**0.547**	0.502	0.500	0.488	0.500	0.488	0.500	0.488	0.500	0.488	0.500	0.488	***0.558***	0.488

Table 5
Specific labels experiment: global results obtained by the individual models and the ensemble without using segmentation-based cropping or data augmentation techniques.

	DenseNet	EfficientNet	Inception	InceptionResNet	Xception	PTC-mode	PTC-lw	CTP
Hamming Loss	0.067	0.107	0.046	0.046	0.046	0.046	0.046	0.046
AUC	0.543	0.547	0.500	0.500	0.500	0.500	0.500	0.558
F1	0.508	0.502	0.488	0.488	0.488	0.488	0.488	0.488

an imbalance between majority and minority classes of 1:172, we can say that the performance of the system is sufficiently high, considering its characteristics.

In the second case study used to validate the proposed methodology, we have grouped the different radiological signs into higher level classes that are more general, as shown in the example of fracture types, Fig. 4. After this grouping, the number of labels passing the minimum 200-sample filter rises to 54 (in the specific labels experiment only 35 labels passed this threshold). Therefore, we now train the system with a larger number of labels, which is closer to the reality of health centres. Regarding the individual models, we can see that the best model is EfficientNet B0 followed by DenseNet, with an AUC of 0.767 and 0.761, respectively. The rest of the models have a value lower than 0.75. Regarding the performance per class of each model, we observe that Xception, EfficientNet and DenseNet perform better in majority classes, while Inception and ResNet perform better in minority classes.

If we look at the results obtained by the ensemble technique, as in the previous case, CTP is the best performer with an AUC of 0.819, which is an improvement of 0.052 over EfficientNet. There are four classes where the ensemble performs as well as the best individual model, but there is no class where the individual models perform better than the ensemble. The number of labels where the ensemble achieves an AUC above 0.85 is slightly lower than in the previous case, 37%, but more than 50% of the classes have an AUC greater than 0.8. This is interesting considering the number of classes (54) and their imbalance. Although the ensemble performs well, it does not perform well for all classes. For example, with the class "Sclerotic bone lesion" it obtains an AUC close to 0.5.

We can observe that in this case the ensemble further improves the individual models as the improvement over the best individual model is now high. The combination of different architectures avoids overfitting and improves the generalisation capacity in a problem where classification is more difficult due to the specificities of the dataset (high number of classes, multilabel, class imbalance). These results demonstrate that this methodology works well on highly imbalanced and multilabel datasets (see Tables 10 and 11).

Table 6
Specific labels experiment: results obtained by training the models with segmentation-based cropping, but without data augmentation. For each label, the individual models with the best performance and the ensembles that outperform all individual models are marked in bold. The best ensemble result is marked in italics.

	# Samples	DenseNet		EfficientNet		Inception		InceptionResnet		Xception		PTC-mode		PTC-lw		CTP	
		AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Normal	34327	0.5	0.374	0.802	0.725	0.453	0.374	**0.819**	0.723	0.5	0.374	0.528	0.444	0.500	0.374	*0.806*	0.374
Copd signs	13419	0.777	0.682	0.785	0.672	0.500	0.457	**0.799**	0.690	0.777	0.682	0.538	0.534	0.648	0.676	***0.825***	0.674
Cardiomegaly	8412	0.900	0.768	0.898	0.774	0.641	0.474	**0.918**	0.767	0.917	0.762	0.596	0.628	0.814	0.792	***0.938***	0.795
Aortic elongation	1399	0.863	0.700	0.874	0.686	0.500	0.479	**0.875**	0.719	0.837	0.705	0.594	0.623	0.767	0.724	***0.898***	0.724
Unchanged	1311	0.612	0.556	**0.625**	0.549	0.500	0.480	0.597	0.549	0.602	0.544	0.506	0.495	0.531	0.539	***0.642***	0.537
Scoliosis	1073	0.823	0.678	0.808	0.690	0.500	0.484	**0.830**	0.702	0.500	0.484	0.591	0.628	0.674	0.711	***0.863***	0.708
Chronic changes	873	0.707	0.537	**0.731**	0.553	0.500	0.487	0.696	0.547	0.695	0.538	0.515	0.518	0.625	0.568	***0.738***	0.568
Costophrenic angle blunting	703	0.810	0.698	0.837	0.691	0.500	0.489	**0.842**	0.655	0.810	0.704	0.558	0.587	0.729	0.713	***0.884***	0.712
Air trapping	663	0.500	0.490	0.671	0.568	0.500	0.490	0.500	0.490	**0.688**	0.553	0.508	0.506	0.500	0.490	***0.705***	0.490
Pleural effusion	658	0.925	0.839	0.942	0.818	0.479	0.046	**0.943**	0.770	0.927	0.838	0.818	0.542	0.901	0.823	*0.942*	0.825
Pneumonia	651	0.759	0.675	0.803	0.671	0.500	0.490	**0.808**	0.655	0.806	0.657	0.572	0.603	0.704	0.691	***0.851***	0.692
Interstitial pattern	594	0.799	0.638	0.795	0.650	0.500	0.491	**0.813**	0.637	0.812	0.615	0.562	0.576	0.714	0.678	***0.858***	0.680
Infiltrates	591	0.733	0.620	0.776	0.635	0.500	0.491	**0.802**	0.597	0.771	0.627	0.563	0.583	0.668	0.639	***0.831***	0.639
Laminar atelectasis	578	0.500	0.491	**0.806**	0.639	0.500	0.491	0.754	0.630	0.745	0.646	0.560	0.587	0.572	0.607	***0.837***	0.607
Vertebral degenerative	575	**0.730**	0.544	0.721	0.540	0.500	0.491	0.725	0.564	0.718	0.533	0.571	0.560	0.620	0.568	***0.771***	0.568
Kyphosis	526	0.796	0.611	**0.813**	0.644	0.500	0.492	0.794	0.628	**0.813**	0.615	0.569	0.585	0.683	0.664	***0.860***	0.664
Apical pleural thickening	469	**0.798**	0.591	0.787	0.573	0.500	0.493	0.775	0.569	0.758	0.575	0.574	0.567	0.701	0.619	***0.838***	0.619
Vascular hilar enlargement	463	0.679	0.562	**0.741**	0.506	0.500	0.493	0.717	0.559	0.715	0.547	0.531	0.533	0.596	0.578	***0.771***	0.582
Fibrotic band	449	0.756	0.568	**0.772**	0.599	0.500	0.493	0.767	0.608	0.758	0.593	0.573	0.585	0.688	0.636	***0.813***	0.638
Nodule	449	0.616	0.557	0.677	0.567	0.500	0.493	**0.688**	0.547	0.626	0.558	0.535	0.545	0.566	0.574	***0.719***	0.572
Calcified granuloma	388	0.741	0.651	0.752	0.641	0.500	0.494	**0.757**	0.622	0.689	0.611	0.578	0.601	0.645	0.654	***0.819***	0.656
Callus rib fracture	360	0.682	0.600	**0.773**	0.594	0.500	0.495	0.500	0.495	0.497	0.495	0.529	0.543	0.500	0.495	***0.799***	0.495
Pacemaker	336	**0.996**	0.948	**0.996**	0.945	0.500	0.495	**0.996**	0.946	**0.996**	0.949	0.741	0.799	0.992	0.951	*0.996*	0.951
Aortic atheromatosis	318	**0.812**	0.538	0.810	0.542	0.500	0.495	0.791	0.567	0.742	0.577	0.544	0.545	0.672	0.605	***0.852***	0.607
Volume loss	294	0.855	0.687	0.862	0.717	0.500	0.496	**0.882**	0.677	0.830	0.691	0.560	0.581	0.762	0.729	***0.910***	0.731
Sternotomy	292	0.991	0.945	0.991	0.939	0.500	0.496	**0.993**	0.872	**0.993**	0.918	0.756	0.814	0.983	0.948	***0.996***	0.948
Bronchiectasis	290	0.673	0.593	**0.726**	0.597	0.500	0.496	0.719	0.587	0.725	0.576	0.541	0.563	0.594	0.613	***0.784***	0.614
Hiatal hernia	287	0.912	0.852	0.920	0.826	0.500	0.496	0.939	0.843	**0.945**	0.726	0.747	0.784	0.877	0.870	***0.962***	0.872
Pseudonodule	275	0.632	0.524	**0.705**	0.547	0.500	0.496	0.536	0.514	0.639	0.540	0.530	0.536	0.540	0.544	***0.718***	0.545
Hemidiaphragm elevation	254	0.902	0.706	0.879	0.697	0.500	0.496	**0.911**	0.696	0.891	0.687	0.749	0.709	0.816	0.751	***0.951***	0.751
Alveolar pattern	248	0.791	0.626	0.834	0.603	0.500	0.496	**0.853**	0.580	0.810	0.604	0.568	0.579	0.715	0.621	***0.887***	0.622
Increased density	239	0.580	0.551	0.586	0.521	0.500	0.496	**0.619**	0.521	0.569	0.526	0.501	0.500	0.533	0.537	***0.634***	0.539
Vertebral anterior compression	214	0.640	0.536	**0.645**	0.530	0.500	0.497	0.644	0.537	0.623	0.517	0.516	0.522	0.524	0.524	***0.702***	0.525
Suture material	210	0.798	0.663	0.791	0.649	0.500	0.497	**0.824**	0.628	0.786	0.665	0.622	0.622	0.742	0.679	***0.833***	0.680
Supra aortic elongation	200	0.697	0.569	0.778	0.564	0.500	0.497	**0.832**	0.561	0.738	0.554	0.579	0.574	0.613	0.577	***0.861***	0.578
Global		0.751	0.633	**0.792**	0.648	0.502	0.475	0.779	0.636	0.750	0.622	0.584	0.586	0.677	0.650	***0.831***	0.651

Table 7
Specific labels experiment: global results obtained by the individual models and the ensemble with preprocessing (segmentation-based cropping) but without data augmentation.

	Densenet201	EfficientNet	Inception	InceptionResnet	Xception	PTC-mode	PTC-lw	CTP
Hamming Loss	0.077	0.079	0.072	0.070	0.077	0.056	0.057	0.057
AUC	0.751	**0.792**	0.502	0.779	0.750	0.584	0.677	**0.831**
F1-score	0.633	0.648	0.475	0.636	0.622	0.586	0.650	0.651

5.3. Visual explanation using heatmaps

As explained in Section 2, the visualisation of multilabel problems is an essential element for this methodology, but it is not a simple problem. Most of the work in this field has deficiencies. Therefore, we have developed a technique that for each label generates a heatmap, an estimated probability, and the ensemble agreement. In Fig. 9, we can see the original X-ray and the heatmaps of the different classes. The areas marked on the radiographs match the radiological signs, and the probabilities are high, with three of the four cases showing agreement between all models.

In the second example, Fig. 10, we can see that the class probabilities are lower than before. The class Atelectasis has an agreement of three models and a low probability (0.583), which means that the physician should be careful with this label. The last example, Fig. 11, belongs to the normal class. In this case, the heat map marks approximately the entire radiograph, as it scans the whole image for radiological signs. The performance of the visualisations is highly dependent on the performance of the model: if the model is better, the visualisations will be more accurate, and the probability and agreement between models will be higher. An advantage of this technique over the state of the art is that we generate a grad-CAM map for each sign that includes the probability generated by the system and the agreement between the models of the ensemble.

6. Discussion

As mentioned throughout the article, the PadChest dataset has a high quality and is really interesting due to the number of classes, which is higher than other multilabel datasets, and the challenge of class imbalance. Although we can find numerous papers using this dataset for medical report generation, it is underutilised in chest X-ray classification problems, which makes the available works for comparison scarce. Moreover, those articles present several problems that complicate an adequate comparison of our work. Therefore, one of our aims is to generate a methodologically correct baseline that allows comparison for future work. For this purpose, we have conducted two experiments: in the first one we have used the specific radiological signs, i.e. the original ones from the dataset, while in the second we have used more generic radiological signs from a tree of terms provided by the authors of the dataset. In Table 12 we can find a summary of

Table 8
Specific labels experiment: results obtained with by training the models with segmentation-based cropping and data augmentation. For each label, the individual models with the best performance and the ensembles that outperform all individual models are marked in bold. The best ensemble result is marked in italics.

	# Samples	Densenet201		EfficientNet		Inception		InceptionResnet		Xception		PTC-mode		PTC-lw		CTP	
		AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Normal	34327	0.820	0.722	0.811	0.716	**0.832**	0.727	0.820	0.709	0.827	0.732	0.725	0.731	0.725	0.731	***0.837***	0.730
Copd signs	13419	**0.823**	0.681	0.785	0.644	0.816	0.678	0.815	0.675	0.800	0.666	0.588	0.610	0.647	0.675	***0.833***	0.672
Cardiomegaly	8412	**0.927**	0.773	0.907	0.749	0.926	0.767	**0.927**	0.777	0.922	0.779	0.746	0.765	0.825	0.789	***0.937***	0.791
Aortic elongation	1399	0.885	0.690	0.846	0.655	0.882	0.676	**0.888**	0.698	**0.888**	0.702	0.690	0.682	0.777	0.705	***0.894***	0.707
Unchanged	1311	0.636	0.553	0.614	0.543	0.638	0.549	**0.642**	0.544	0.636	0.547	0.531	0.537	0.545	0.551	*0.641*	0.551
Scoliosis	1073	0.759	0.636	0.712	0.598	0.745	0.605	0.732	0.602	**0.774**	0.661	0.630	0.637	0.671	0.659	***0.793***	0.664
Chronic changes	873	0.759	0.518	0.720	0.549	**0.768**	0.546	0.762	0.519	0.752	0.533	0.621	0.556	0.684	0.545	***0.772***	0.547
Costophrenic angle blunting	703	**0.862**	0.674	0.845	0.674	0.832	0.662	0.831	0.665	0.855	0.663	0.685	0.676	0.739	0.693	***0.877***	0.693
Air trapping	663	**0.692**	0.557	0.687	0.560	0.515	0.490	0.469	0.490	0.506	0.490	0.525	0.532	0.500	0.490	***0.704***	0.490
Pleural effusion	658	**0.959**	0.827	0.951	0.811	0.956	0.823	0.955	0.830	0.945	0.816	0.862	0.822	0.886	0.839	***0.967***	0.840
Pneumonia	651	0.815	0.671	**0.821**	0.660	0.810	0.663	**0.821**	0.672	**0.821**	0.668	0.681	0.668	0.703	0.687	***0.850***	0.687
Interstitial pattern	594	0.834	0.625	0.828	0.636	**0.846**	0.651	0.843	0.616	0.830	0.613	0.727	0.651	0.743	0.650	***0.858***	0.651
Infiltrates	591	0.812	0.629	0.803	0.617	0.803	0.626	**0.815**	0.633	0.808	0.639	0.649	0.635	0.662	0.644	***0.840***	0.646
Laminar atelectasis	578	**0.843**	0.670	0.812	0.637	0.827	0.643	0.827	0.654	0.833	0.654	0.666	0.658	0.690	0.677	***0.858***	0.678
Vertebral degenerative changes	575	0.779	0.545	0.730	0.544	0.774	0.546	**0.785**	0.518	0.779	0.547	0.627	0.560	0.670	0.557	***0.797***	0.556
Kyphosis	526	**0.867**	0.640	0.834	0.589	0.845	0.587	0.849	0.609	0.839	0.625	0.691	0.617	0.736	0.639	***0.870***	0.640
Apical pleural thickening	469	**0.808**	0.553	0.801	0.568	0.789	0.573	0.509	0.493	0.500	0.493	0.592	0.574	0.661	0.619	***0.830***	0.621
Vascular hilar enlargement	463	**0.746**	0.549	0.742	0.568	0.769	0.522	0.755	0.544	0.745	0.515	0.618	0.556	0.651	0.562	***0.783***	0.563
Fibrotic band	449	**0.831**	0.583	0.809	0.600	0.813	0.614	0.575	0.493	0.806	0.611	0.641	0.604	0.716	0.658	***0.848***	0.659
Nodule	449	**0.706**	0.578	0.675	0.554	0.561	0.493	0.574	0.493	0.551	0.493	0.518	0.526	0.500	0.493	*0.704*	0.493
Calcified granuloma	388	**0.808**	0.653	0.802	0.649	0.542	0.494	0.554	0.494	0.496	0.494	0.572	0.593	0.500	0.494	***0.833***	0.494
Callus rib fracture	360	0.717	0.606	**0.765**	0.557	0.614	0.495	0.609	0.495	0.571	0.495	0.550	0.549	0.500	0.495	***0.787***	0.495
Pacemaker	336	0.993	0.927	**0.997**	0.942	0.996	0.919	0.996	0.931	0.984	0.926	0.984	0.930	0.993	0.946	***0.997***	0.946
Aortic atheromatosis	318	0.856	0.521	0.847	0.516	0.862	0.550	0.852	0.541	**0.871**	0.559	0.739	0.556	0.786	0.558	***0.885***	0.557
Volume loss	294	**0.917**	0.693	0.902	0.657	0.904	0.636	0.896	0.672	0.902	0.640	0.789	0.670	0.809	0.693	***0.928***	0.697
Sternotomy	292	0.992	0.898	0.987	0.920	0.990	0.926	**0.995**	0.936	0.992	0.865	0.961	0.912	0.984	0.941	***0.997***	0.941
Bronchiectasis	290	0.801	0.549	0.775	0.561	0.796	0.578	**0.805**	0.562	0.794	0.573	0.682	0.580	0.690	0.587	***0.820***	0.588
Hiatal hernia	287	0.939	0.856	0.941	0.697	0.947	0.824	**0.964**	0.801	0.947	0.851	0.871	0.786	0.876	0.867	***0.967***	0.867
Pseudonodule	275	0.612	0.496	**0.670**	0.550	0.598	0.496	0.589	0.496	0.545	0.496	0.536	0.543	0.500	0.496	***0.672***	0.496
Hemidiaphragm elevation	254	0.893	0.667	0.882	0.679	**0.915**	0.670	0.894	0.702	0.898	0.684	0.759	0.692	0.784	0.725	***0.934***	0.724
Alveolar pattern	248	0.876	0.627	0.877	0.589	0.885	0.607	**0.895**	0.606	0.871	0.594	0.748	0.612	0.774	0.623	***0.911***	0.622
Increased density	239	0.643	0.541	0.641	0.509	0.651	0.534	0.633	0.526	**0.668**	0.535	0.545	0.531	0.547	0.544	***0.673***	0.547
Vertebral anterior compression	214	**0.749**	0.532	0.696	0.524	0.736	0.517	0.743	0.527	0.734	0.535	0.573	0.530	0.597	0.540	***0.752***	0.539
Suture material	210	0.819	0.652	0.820	0.663	0.818	0.639	0.811	0.662	**0.822**	0.612	0.759	0.663	0.768	0.685	***0.847***	0.684
Supra aortic elongation	200	0.865	0.576	0.821	0.541	0.880	0.546	**0.882**	0.563	0.857	0.562	0.628	0.558	0.681	0.576	***0.894***	0.575
Global		**0.818**	0.642	0.804	0.629	0.797	0.625	0.780	0.621	0.782	0.625	0.677	0.637	0.701	0.647	***0.840***	0.647

Table 9
Specific labels experiment: global results obtained from the individual models and the ensembles.

	Densenet201	EfficientNet	Inception	InceptionResnet	Xception	PTC-mode	PTC-lw	CTP
Hamming Loss	0.082	0.078	0.081	0.077	0.074	0.063	0.065	0.065
AUC	0.818	0.804	0.797	0.780	0.782	0.677	0.701	0.840
F1	0.642	0.629	0.625	0.621	0.625	0.637	0.647	0.647

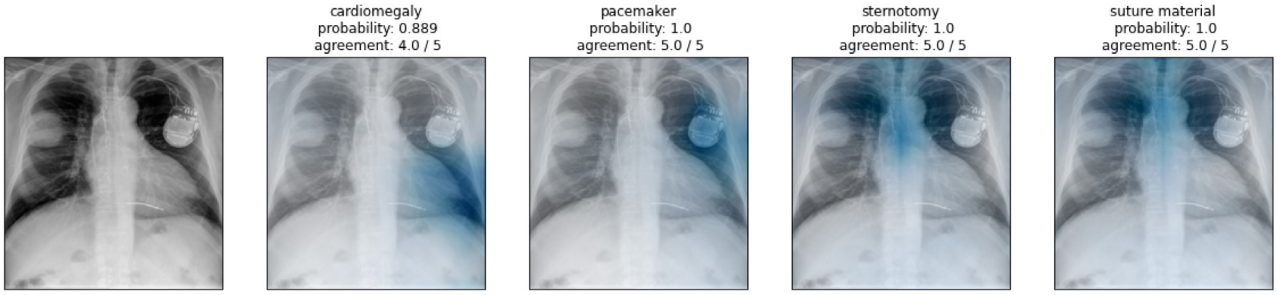

Fig. 9. First visualisation example. The heatmaps of four radiological signs detected (cardiomegaly, pacemaker, sternotomy and suture material) are shown. The title shows the label, the probability estimated by the ensemble, and the agreement between the models of the ensemble. The areas of interest for classification are marked in blue.

the different published systems and their global and class specific performance. First of all, it is interesting to note how most of the papers have selected different labels to perform the classification, and all papers, except [61], select a low number of total classes compared to the number of classes available.

In the case of Rimeika et al. [87], the publication does not show how the two models have been built; it does not provide information on the architecture, the other dataset used, or the criteria for selecting the classes from PadChest dataset, so there is no possibility to replicate these models, and therefore we cannot use it for comparison. In Pooch et al. [62], the PadChest classes have been adapted to match the classes of other multilabel datasets such as ChestX-ray14 and CheXpert. For example, regardless of the fact that the class "Lesion" does not exist in two of the datasets, they generated this class using the ChestX-ray14 labels "Nodules" and "Masses". However, PadChest was processed in

Table 10
General labels experiment: results obtained by training the models with segmentation-based cropping and data augmentation. For each label, the individual models with the best performance and the ensembles that outperform all individual models are marked in bold. The best ensemble result is marked in italics.

	# Samples	Densenet201		EfficientNet		Inception		InceptionResnet		Xception		PTC-mode		PTC-lw		CTP	
		AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Normal	34327	0.735	0.685	0.707	0.652	**0.750**	0.691	0.723	0.658	0.732	0.674	0.702	0.693	0.690	0.691	***0.770***	0.691
Copd signs	13419	0.771	0.629	0.761	0.649	0.771	0.618	**0.793**	0.665	0.779	0.645	0.596	0.621	0.615	0.645	***0.816***	0.640
Cardiomegaly	8120	0.899	0.736	**0.904**	0.746	0.890	0.723	0.892	0.751	0.898	0.741	0.766	0.747	0.816	0.760	***0.923***	0.762
Thoracic cage deformation	7778	0.706	0.603	**0.728**	0.627	0.500	0.478	0.675	0.595	0.708	0.609	0.577	0.586	0.601	0.612	***0.745***	0.614
Aortic elongation	7436	0.858	0.691	0.853	0.690	0.842	0.661	**0.866**	0.683	**0.866**	0.687	0.701	0.690	0.758	0.697	***0.886***	0.700
Infiltrates	6706	0.794	0.686	**0.802**	0.664	0.791	0.663	0.797	0.668	0.794	0.663	0.703	0.676	0.721	0.690	***0.827***	0.692
Unchanged	6487	0.630	0.538	**0.636**	0.552	0.618	0.543	0.633	0.545	0.631	0.552	0.536	0.543	0.536	0.546	***0.652***	0.547
Chronic changes	4312	**0.759**	0.548	0.754	0.542	0.752	0.525	0.734	0.520	0.740	0.522	0.656	0.567	0.685	0.543	***0.768***	0.545
Surgery	3928	0.813	0.730	0.815	0.766	0.750	0.722	0.766	0.713	**0.829**	0.724	0.726	0.739	0.739	0.762	***0.845***	0.765
Atelectasis	3565	**0.798**	0.628	0.756	0.636	0.698	0.570	0.729	0.596	0.759	0.628	0.587	0.598	0.647	0.632	***0.804***	0.630
Costophrenic angle blunting	3306	**0.845**	0.655	0.807	0.638	0.758	0.604	0.784	0.638	0.828	0.652	0.660	0.634	0.700	0.656	***0.864***	0.658
Calcified densities	3253	0.719	0.638	**0.751**	0.639	0.500	0.491	0.500	0.491	0.500	0.491	0.520	0.527	0.500	0.491	***0.764***	0.491
Vertebral degenerative changes	3203	**0.744**	0.502	0.726	0.528	0.676	0.497	0.733	0.487	0.730	0.512	0.643	0.532	0.664	0.514	***0.751***	0.514
Hilar enlargement	3162	**0.755**	0.549	0.732	0.544	0.699	0.551	0.738	0.533	0.731	0.538	0.618	0.562	0.649	0.560	***0.765***	0.565
Pleural thickening	3010	0.753	0.586	**0.773**	0.585	0.737	0.572	0.743	0.525	0.763	0.562	0.651	0.587	0.671	0.586	***0.790***	0.587
Mediastinal enlargement	2813	0.795	0.643	0.798	0.668	0.774	0.688	0.778	0.675	**0.822**	0.657	0.705	0.689	0.710	0.697	***0.841***	0.700
Air trapping	2765	0.654	0.528	0.669	0.536	0.500	0.492	0.665	0.495	**0.672**	0.536	0.520	0.523	0.602	0.544	***0.692***	0.546
Fracture	2529	**0.749**	0.663	0.725	0.599	0.5	0.493	0.640	0.507	0.732	0.611	0.574	0.590	0.579	0.615	***0.792***	0.617
Pleural effusion	2436	**0.942**	0.738	0.927	0.782	0.930	0.720	0.935	0.735	0.937	0.771	0.878	0.762	0.900	0.775	***0.956***	0.775
Granuloma	2306	0.500	0.493	**0.777**	0.652	0.500	0.493	0.500	0.493	0.500	0.493	0.513	0.519	0.500	0.493	***0.777***	0.493
Nodule	1936	0.653	0.583	**0.679**	0.571	0.617	0.542	0.643	0.547	0.622	0.575	0.560	0.569	0.558	0.575	***0.707***	0.573
Fibrotic band	1781	0.738	0.530	**0.747**	0.531	0.712	0.522	0.500	0.495	0.727	0.519	0.606	0.546	0.654	0.556	***0.770***	0.556
Electrical device	1772	0.992	0.959	0.992	0.913	0.992	0.889	**0.994**	0.871	0.992	0.929	0.990	0.935	0.992	0.942	***0.997***	0.942
Pneumonia	1652	0.804	0.594	0.790	0.567	0.804	0.549	**0.813**	0.577	0.799	0.599	0.712	0.602	0.728	0.606	***0.854***	0.607
Aortic atheromatosis	1581	0.834	0.502	0.830	0.540	0.813	0.477	0.840	0.524	**0.843**	0.519	0.713	0.554	0.769	0.522	***0.866***	0.523
Pseudonodule	1451	0.693	0.561	**0.727**	0.553	0.500	0.496	0.500	0.496	0.708	0.562	0.557	0.555	0.576	0.589	***0.759***	0.593
Bronchiectasis	1430	0.795	0.544	0.776	0.571	0.789	0.548	**0.814**	0.539	0.779	0.561	0.639	0.575	0.696	0.573	***0.833***	0.574
Hiatal hernia	1362	0.916	0.813	0.892	0.796	0.906	0.773	0.918	0.804	**0.927**	0.788	0.857	0.810	0.889	0.852	***0.959***	0.852
Hemidiaphragm elevation	1231	0.814	0.651	**0.841**	0.680	**0.841**	0.596	0.823	0.649	0.811	0.645	0.733	0.670	0.746	0.683	***0.890***	0.683
Increased density	1133	0.633	0.497	**0.640**	0.524	0.596	0.492	0.609	0.509	0.606	0.511	0.539	0.516	0.533	0.514	***0.661***	0.515
Diaphragmatic eventration	757	0.500	0.498	**0.775**	0.586	0.500	0.498	0.500	0.498	0.500	0.498	0.525	0.534	0.500	0.498	***0.775***	0.498
Volume loss	684	0.802	0.542	0.776	0.580	0.809	0.531	**0.814**	0.513	0.776	0.560	0.728	0.561	0.761	0.564	***0.865***	0.564
Adenopathy	659	0.500	0.498	**0.697**	0.538	0.500	0.498	0.548	0.520	0.583	0.543	0.500	0.498	0.521	0.528	***0.715***	0.528
Bronchovascular markings	602	0.712	0.570	0.738	0.537	**0.777**	0.576	0.765	0.545	0.704	0.585	0.685	0.592	0.703	0.585	***0.802***	0.584
Mass	574	0.707	0.621	0.715	0.608	0.744	0.570	0.732	0.574	**0.746**	0.616	0.700	0.615	0.707	0.641	***0.806***	0.641
Artificial heart valve	562	0.969	0.658	0.953	0.730	0.975	0.696	**0.977**	0.727	0.972	0.713	0.941	0.736	0.968	0.730	***0.981***	0.731
Catheter	545	0.871	0.740	0.874	0.721	0.866	0.673	**0.878**	0.639	0.861	0.717	0.799	0.724	0.848	0.773	***0.905***	0.773
Suboptimal study	544	0.743	0.524	0.693	0.510	**0.754**	0.522	0.697	0.531	0.727	0.506	0.666	0.526	0.681	0.539	***0.784***	0.540
Pulmonary fibrosis	523	0.850	0.584	0.834	0.587	0.837	0.551	**0.864**	0.577	0.862	0.565	0.795	0.577	0.810	0.591	***0.892***	0.591
Heart insufficiency	520	0.875	0.541	0.877	0.555	0.896	0.546	**0.884**	0.538	0.870	0.547	0.819	0.553	0.856	0.551	***0.920***	0.551
Hypoexpansion	476	0.838	0.541	0.745	0.545	**0.846**	0.534	0.768	0.571	0.500	0.499	0.651	0.556	0.677	0.571	***0.900***	0.573
Gynecomastia	437	0.852	0.527	0.810	0.552	0.852	0.501	**0.858**	0.507	0.806	0.550	0.772	0.540	0.825	0.554	***0.917***	0.555
Emphysema	410	0.780	0.508	0.715	0.520	0.801	0.512	**0.809**	0.506	0.724	0.521	0.684	0.529	0.732	0.524	***0.862***	0.525
Sclerotic bone lesion	352	**0.506**	0.511	0.500	0.499	0.500	0.499	0.500	0.499	0.500	0.499	0.500	0.499	0.500	0.499	***0.506***	0.499
Fissure thickening	336	0.816	0.533	0.802	0.573	0.819	0.518	**0.842**	0.526	0.798	0.539	0.746	0.547	0.806	0.557	***0.891***	0.558
Hilar congestion	318	0.785	0.503	0.798	0.519	0.790	0.514	**0.827**	0.519	0.808	0.520	0.734	0.526	0.756	0.523	***0.896***	0.522
Osteopenia	318	0.659	0.508	0.688	0.507	0.659	0.483	0.695	0.466	**0.701**	0.497	0.611	0.500	0.647	0.500	***0.752***	0.500
Tuberculosis	299	0.852	0.534	0.861	0.567	**0.869**	0.561	0.824	0.559	0.805	0.597	0.760	0.577	0.848	0.592	***0.909***	0.592
Bullas	290	**0.746**	0.520	0.685	0.532	0.739	0.524	0.715	0.512	0.651	0.549	0.667	0.543	0.714	0.547	***0.777***	0.547
Hyperinflated lung	272	0.715	0.506	0.630	0.502	**0.719**	0.504	0.645	0.485	0.659	0.501	0.649	0.513	0.658	0.512	***0.728***	0.512
Cavitation	243	0.780	0.556	0.834	0.575	**0.856**	0.546	0.789	0.539	0.823	0.590	0.679	0.555	0.823	0.585	***0.934***	0.585
Mediastinic lipomatosis	212	0.648	0.499	**0.654**	0.551	0.5	0.499	0.500	0.499	0.500	0.499	0.520	0.514	0.500	0.499	***0.681***	0.499
Pneumothorax	210	0.705	0.572	0.717	0.530	0.717	0.540	**0.721**	0.518	0.620	0.592	0.596	0.546	0.630	0.572	***0.847***	0.573
Vascular redistribution	204	**0.774**	0.499	0.752	0.526	0.705	0.508	0.694	0.516	0.667	0.507	0.635	0.514	0.676	0.519	***0.837***	0.519
Global		0.761	0.589	**0.767**	0.600	0.732	0.566	0.739	0.572	0.739	0.589	0.669	0.594	0.696	0.601	***0.819***	0.602

Table 11
General labels experiment: global results obtained by the individual models and the ensembles.

	Densenet201	EfficientNet	Inception	InceptionResnet	Xception	PTC-mode	PTC-lw	CTP
Hamming Loss	0.070	0.065	0.070	0.075	0.065	0.052	0.057	0.056
AUC	0.761	0.767	0.732	0.739	0.739	0.669	0.696	0.819
F1-score	0.589	0.600	0.566	0.572	0.589	0.594	0.601	0.602

that paper by unifying all classes related to Atelectasis, without providing any medical explanation for this decision, so the labels do not match ours, containing only 8 out of 174 classes available. Therefore, it cannot be compared with our methodology. In the case of Hashir et al. [61], they first select a single sample from each patient and the authors have used 32 different labels, which is not in line with expectations, since using a lower threshold than ours there should have a larger number of labels. In this case, we can compare the overall AUC of the system. These authors achieve an AUC of 0.800 using 32 labels while we obtain 0.8397 using 35 specific labels. Therefore, we have achieved a better AUC than in Hashir et al. [61]. Because of the above reasons, comparing our methodology with the state of the art is really difficult, and therefore one of the aims of this paper is to create a baseline

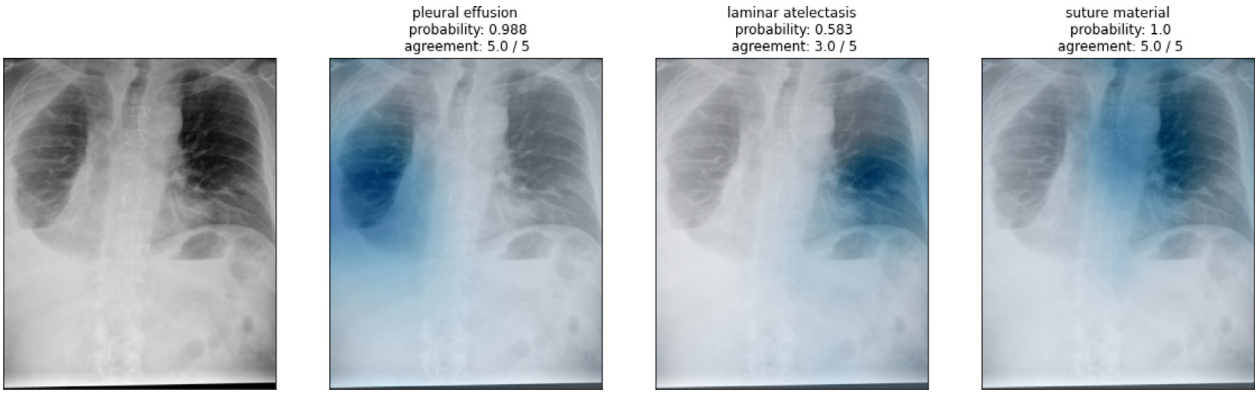

Fig. 10. Second visualisation example. The heatmaps of three radiological signs detected (pleural effusion, laminar atelectasis and suture material) are shown. The title shows the label, the probability estimated by the ensemble, and the agreement between the models of the ensemble. The areas of interest for classification are marked in blue.

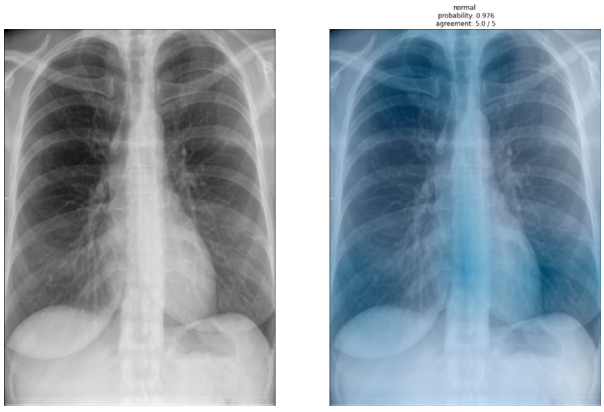

Fig. 11. Third visualisation example. The sample belongs to the normal class. The title shows the label, the probability estimated by the ensemble, and the agreement between the models of the ensemble. The areas of interest for classification are marked in blue.

Table 12
Comparative table of the different state-of-the-art models, their global and class performance.

	Rimeika G. et al. [87]		Pooch, E. H. [62]
	model1	model2	
cardiomegaly	90.36%	91.94%	90.75%
nodule	74.97%	71.42%	–
normal	–	–	87.10%
pleural effusion	95.42%	94.93%	–
pneumonia	–	–	79.90%
lobar collapse	88.86%	86.39%	–
edema	95.35%	96.05%	91.07%
subcutaneous emphysema	98.52%	93.79%	–
consolidation	87.39%	85.50%	86.07%
pneumothorax	89.95%	88.19%	82.76%
tuberculosis	92.62%	92.40%	–
Lymphadenopathy	77.11%	75.81%	–
linear atelectasis	84.16%	78.26%	76.41%
lymph node calficiation	82.64%	72.69%	–
congestion	85.39%	87.29%	–
Widened mediastinum	75.02%	77.50%	–
mass	86.90%	82.29%	–
lesion	–	–	69.75%
Global	*86.98%*	*84.97%*	*82.98%*

to facilitate the comparison of future work with this dataset. If we look at the overall AUC of the published models trained with PadChest and compare them with ours, we see that we only outperform two models, but we use a much higher number of classes. Therefore, we can see that our system performs well and that although it works with a much larger number of labels, it outperforms some of the published models.

7. Conclusions and future work

This paper proposes a Deep Learning methodology for classification tasks with imbalanced multilabel datasets. We have built with this methodology an ensemble of five state-of-the-art architectures: DenseNet-201, EfficientNet B0, Inception, InceptionResNet and Xception. We have used weighted crossentropy with logit loss to alleviate data imbalance and developed a new technique for generating heatmaps in multilabel classification problems.

The results of our experiments are promising. First, in contrast to state-of-the-art papers, we have established a methodologically sound baseline for future work, regardless of whether specific or general labels are used. It will also allow us to analyse the performance of these models when the number of labels varies. Our system obtains high AUC values for the number of classes used. In the case of specific labels, high performance is achieved with an AUC of 0.84. In the case of general labels, we obtain an AUC of 0.819. This value may be due to the fact that the general classification has more classes and each of them is composed of different radiological signs. Thus, the variability is high and it is more difficult to classify. The results of the visualisation technique show a great potential, as it allows a view of the whole radiograph that differentiates the different pathological signs. This technique generates a report that includes the visualisation of the heatmap, the probability produced by the system and the agreement between the ensemble models.

There are several ways to improve our methodology. First, other strategies can be used to alleviate data imbalance, such as adding new samples to the dataset. This can be done either by obtaining new images from other datasets such as CheXpert, ChestX-ray14, or other single disease datasets, or by creating them with generative adversarial networks (GANs). Another way to improve the performance of the proposed system is to use different X-ray views of each sample.

In our proposal, we used segmentation techniques to force the model to pay more attention to the most relevant areas. However, different techniques have recently been developed for this same purpose. For example, [88] used soft and hard attention mechanisms to prevent the model from focusing on areas that are not relevant to the problem. Another way to remove non-interesting areas is the application of semi-supervised learning methods to locate and distinguish different anatomical regions [89]. Based on

these recent advances, it would be interesting to study whether including them in our model improves its results.

To improve the visualisation technique, we can extend the displayed information by including a heatmap that shows the standard deviation of the visualisation of the ensemble [90]. This would help medical staff to know in which areas of the heatmap there is more uncertainty. Another line of work we would like to explore is the generation of a system that returns general and specific labels. In addition to a combination with report generation techniques, doctors would receive a report explaining the different radiological signs and a visual interpretation of these signs. This could be done using cascade models, which first classify the most general labels and later classify the subcategories. This would allow to include minority classes, or at least part of them.

Another possible improvement would be to retrain the system using feedback from experts in the field on the system's predictions and heatmaps. This poses multiple challenges in practice, mainly due to the need to implement close collaboration between specialised diagnostic models and medical staff, who may lack background and expertise to rely on the models' results. On the positive side, semi-supervised Deep Learning techniques are emerging lately and are yielding results that are unprecedented in the state of the art. For instance, Avilés-Rivero et al. [91] have developed a semi-supervised graph-based framework for classifying lung diseases (COVID-19, pneumonia and healthy). Such frameworks have been identified as promising for supporting the construction of human-in-the-loop models in medical applications [30], hence future efforts will be devoted in this direction.

CRediT authorship contribution statement

Helena Liz: Conceptualization, Visualization, Writing, Methodology. **Javier Huertas-Tato:** Conceptualization, Visualization, Validation, Methodology, Writing. **Manuel Sánchez-Montañés:** Conceptualization, Writing – review & editing. **Javier Del Ser:** Conceptualization, Writing – review & editing. **David Camacho:** Conceptualization, Writing, Resources, Funding acquisition, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

This work has been funded by Grant PLEC2021-007681 (XAI-DisInfodemics) and PID2020-117263GB-100 (FightDIS) funded by MCIN/AEI/ 10.13039/501100011033 and, as appropriate, by "ERDF A way of making Europe", by the "European Union NextGenerationEU/PRTR", by the research project CIVIC: Intelligent characterisation of the veracity of the information related to COVID-19, granted by BBVA FOUNDATION GRANTS FOR SCIENTIFIC RESEARCH TEAMS SARS-CoV-2 and COVID-19, by European Comission under IBERIFIER - Iberian Digital Media Research and Fact-Checking Hub (2020-EU-IA-0252), by "Convenio Plurianual with the Universidad Politécnica de Madrid in the actuation line of Programa de Excelencia para el Profesorado Universitario", and by Comunidad Autónoma de Madrid under S2018/TCS-4566 (CYNAMON) grant. M. Sánchez-Montañés has been supported by grants PID2021-127946OB-I00 and PID2021-122347NB-I00 (funded by MCIN/AEI/ 10.13039/501100011033 and ERDF - "A way of making Europe") and Comunidad Autónoma de Madrid, Spain (S2017/BMD-3688 MULTI-TARGET&VIEW-CM grant). J. Del Ser thanks the financial support of the Spanish Centro para el Desarrollo Tecnológico Industrial (CDTI, Ministry of Science and Innovation) through the "Red Cervera" Programme (AI4ES project), as well as the support of the Basque Government (consolidated research group MATHMODE, ref. IT1456-22)

References

[1] E. Moustaka, T.C. Constantinidis, Sources and effects of work-related stress in nursing, Health Sci. J. 4 (4) (2010) 210.
[2] S. Domínguez-Rodríguez, H. Liz, A. Panizo, Á. Ballesteros, R. Dagan, D. Greenberg, L. Gutiérrez, P. Rojo, E. Otheo, J.C. Galán, et al., Testing the performance, adequacy, and applicability of an artificial intelligent model for pediatric pneumonia diagnosis, 2022.
[3] N. Shaw, M. Hendry, O. Eden, Inter-observer variation in interpretation of chest X-rays, Scott. Med. J. 35 (5) (1990) 140–141.
[4] K.B. Ahmed, G.M. Goldgof, R. Paul, D.B. Goldgof, L.O. Hall, Discovery of a generalization gap of Convolutional Neural Networks on COVID-19 X-rays classification, Ieee Access 9 (2021) 72970–72979.
[5] Y. LeCun, Y. Bengio, et al., Convolutional networks for images, speech, and time series, in: The Handbook of Brain Theory and Neural Networks, Vol. 3361, no. 10, 1995, p. 1995.
[6] T. Kontzer, Deep learning drops error rate for breast cancer diagnoses by 85%, 2016, URL: https://blogs.nvidia.com/blog/2016/09/19/deep-learning-breast-cancer-diagnosis/.
[7] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[8] T. Agrawal, P. Choudhary, EfficientUNet: Modified encoder-decoder architecture for the lung segmentation in chest X-ray images, Expert Syst. (2022) e13012.
[9] I.M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, A. Saalbach, Comparison of deep learning approaches for multi-label chest X-ray classification, Sci. Rep. 9 (1) (2019) 1–10.
[10] J.-Y. Park, Y. Hwang, D. Lee, J.-H. Kim, MarsNet: Multi-label classification network for images of various sizes, IEEE Access 8 (2020) 21832–21846.
[11] A. Bustos, A. Pertusa, J.-M. Salinas, M. de la Iglesia-Vayá, Padchest: A large chest X-ray image dataset with multi-label annotated reports, Med. Image Anal. 66 (2020) 101797.
[12] I. Al-Badarneh, M. Habib, I. Aljarah, H. Faris, Neuro-evolutionary models for imbalanced classification problems, J. King Saud Univ.-Comput. Inf. Sci. (2020).
[13] E. Lin, Q. Chen, X. Qi, Deep reinforcement learning for imbalanced classification, Appl. Intell. 50 (8) (2020) 2488–2502.
[14] T.N. Mundhenk, B.Y. Chen, G. Friedland, Efficient saliency maps for explainable AI, 2019, arXiv preprint arXiv:1911.11293.
[15] A. Holzinger, M. Dehmer, F. Emmert-Streib, R. Cucchiara, I. Augenstein, J. Del Ser, W. Samek, I. Jurisica, N. Díaz-Rodríguez, Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence, Inf. Fusion 79 (2022) 263–278.
[16] R.K. Singh, R. Pandey, R.N. Babu, Covidscreen: explainable deep learning framework for differential diagnosis of COVID-19 using chest X-rays, Neural Comput. Appl. 33 (14) (2021) 8871–8892.
[17] F. Piccialli, V. Di Somma, F. Giampaolo, S. Cuomo, G. Fortino, A survey on deep learning in medicine: Why, how and when? Inf. Fusion 66 (2021) 111–137.
[18] S.W. Baalman, F.E. Schroevers, A.J. Oakley, T.F. Brouwer, W. van der Stuijt, H. Bleijendaal, L.A. Ramos, R.R. Lopes, H.A. Marquering, R.E. Knops, et al., A morphology based deep learning model for atrial fibrillation detection using single cycle electrocardiographic samples, Int. J. Cardiol. 316 (2020) 130–136.
[19] E.A. Chung, M.E. Benalcázar, Real-time hand gesture recognition model using deep learning techniques and EMG signals, in: 2019 27th European Signal Processing Conference, EUSIPCO, IEEE, 2019, pp. 1–5.
[20] H. Li, X. Wang, C. Liu, Q. Zeng, Y. Zheng, X. Chu, L. Yao, J. Wang, Y. Jiao, C. Karmakar, A fusion framework based on multi-domain features and deep learning features of phonocardiogram for coronary artery disease detection, Comput. Biol. Med. 120 (2020) 103733.
[21] J. Pan, Y. Zi, J. Chen, Z. Zhou, B. Wang, LiftingNet: A novel deep learning network with layerwise feature learning from noisy mechanical data for fault classification, IEEE Trans. Ind. Electron. 65 (6) (2017) 4973–4982.
[22] Y.-S. Su, T.-J. Ding, M.-Y. Chen, Deep learning methods in internet of medical things for valvular heart disease screening system, IEEE Internet Things J. 8 (23) (2021) 16921–16932.

[23] F.N. Abdullah, M.N. Fauzan, N. Riza, Multiple linear regression and deep learning in body temperature detection and mask detection, IT J. Res. Dev. (2022) 109–121.
[24] M. Jost, D.A. Santos, R.A. Saunders, M.A. Horlbeck, J.S. Hawkins, S.M. Scaria, T.M. Norman, J.A. Hussmann, C.R. Liem, C.A. Gross, et al., Titrating gene expression using libraries of systematically attenuated CRISPR guide RNAs, Nature Biotechnol. 38 (3) (2020) 355–364.
[25] S. Li, K. Yu, D. Wang, Q. Zhang, Z.-X. Liu, L. Zhao, H. Cheng, Deep learning based prediction of species-specific protein S-glutathionylation sites, Biochim. Biophys. Acta (BBA)-Proteins Proteomics 1868 (7) (2020) 140422.
[26] G. Zampieri, S. Vijayakumar, E. Yaneske, C. Angione, Machine and deep learning meet genome-scale metabolic modeling, PLoS Comput. Biol. 15 (7) (2019) e1007084.
[27] M.L. Welch, C. McIntosh, A. McNiven, S.H. Huang, B.-B. Zhang, L. Wee, A. Traverso, B. O'Sullivan, F. Hoebers, A. Dekker, et al., User-controlled pipelines for feature integration and head and neck radiation therapy outcome predictions, Phys. Med. 70 (2020) 145–152.
[28] Z. Xu, J. Chou, X.S. Zhang, Y. Luo, T. Isakova, P. Adekkanattu, J.S. Ancker, G. Jiang, R.C. Kiefer, J.A. Pacheco, et al., Identifying sub-phenotypes of acute kidney injury using structured and unstructured electronic health record data with memory networks, J. Biomed. Inform. 102 (2020) 103361.
[29] K. Rough, A.M. Dai, K. Zhang, Y. Xue, L.M. Vardoulakis, C. Cui, A.J. Butte, M.D. Howell, A. Rajkomar, Predicting inpatient medication orders from electronic health record data, Clin. Pharmacol. Therapeutics 108 (1) (2020) 145–154.
[30] I. Ahmed, D. Camacho, G. Jeon, F. Piccialli, Internet of health things driven deep learning-based system for non-invasive patient discomfort detection using time frame rules and pairwise keypoints distance feature, Sustainable Cities Soc. 79 (2022) 103672.
[31] D. Arefan, A.A. Mohamed, W.A. Berg, M.L. Zuley, J.H. Sumkin, S. Wu, Deep learning modeling using normal mammograms for predicting breast cancer risk, Med. Phys. 47 (1) (2020) 110–118.
[32] M. Byra, M. Wu, X. Zhang, H. Jang, Y.-J. Ma, E.Y. Chang, S. Shah, J. Du, Knee menisci segmentation and relaxometry of 3D ultrashort echo time cones MR imaging using attention U-Net with transfer learning, Magn. Reson. Med. 83 (3) (2020) 1109–1122.
[33] S. Saha, A. Pagnozzi, P. Bourgeat, J.M. George, D. Bradford, P.B. Colditz, R.N. Boyd, S.E. Rose, J. Fripp, K. Pannek, Predicting motor outcome in preterm infants from very early brain diffusion MRI using a deep learning Convolutional Neural Network (CNN) model, Neuroimage 215 (2020) 116807.
[34] M. Saminathan, M. Ramachandran, A. Kumar, K. Rajkumar, A. Khanna, P. Singh, A study on specific learning algorithms pertaining to classify lung cancer disease, Expert Syst. 39 (3) (2022) e12797.
[35] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.
[36] S. Kazemifar, A.M. Barragán Montero, K. Souris, S.T. Rivas, R. Timmerman, Y.K. Park, S. Jiang, X. Geets, E. Sterpin, A. Owrangi, Dosimetric evaluation of synthetic CT generated with GANs for MRI-only proton therapy treatment planning of brain tumors, J. Appl. Clin. Med. Phys. 21 (5) (2020) 76–86.
[37] G. Liu, T.-M.H. Hsu, M. McDermott, W. Boag, W.-H. Weng, P. Szolovits, M. Ghassemi, Clinically accurate chest X-ray report generation, in: Machine Learning for Healthcare Conference, PMLR, 2019, pp. 249–269.
[38] F. Piccialli, F. Giampaolo, E. Prezioso, D. Camacho, G. Acampora, Artificial intelligence and healthcare: Forecasting of medical bookings through multi-source time-series fusion, Inf. Fusion 74 (2021) 1–16.
[39] T. Nemoto, N. Futakami, M. Yagi, A. Kumabe, A. Takeda, E. Kunieda, N. Shigematsu, Efficacy evaluation of 2D, 3D U-Net semantic segmentation and atlas-based segmentation of normal lungs excluding the trachea and main bronchi, J. Radiat. Res. 61 (2) (2020) 257–264.
[40] D.C. Benz, G. Benetos, G. Rampidis, E. Von Felten, A. Bakula, A. Sustar, K. Kudura, M. Messerli, T.A. Fuchs, C. Gebhard, et al., Validation of deep-learning image reconstruction for coronary computed tomography angiography: Impact on noise, image quality and diagnostic accuracy, J. Cardiovasc. Comput. Tomography 14 (5) (2020) 444–451.
[41] T.D. Pham, Classification of COVID-19 chest X-rays with deep learning: New models or fine tuning? Health Inf. Sci. Syst. 9 (1) (2021) 1–11.
[42] F. Ahmad, A. Farooq, M.U. Ghani, Deep ensemble model for classification of novel coronavirus in chest X-ray images, Comput. Intell. Neurosci. 2021 (2021).
[43] D. Avola, A. Bacciu, L. Cinque, A. Fagioli, M.R. Marini, R. Taiello, Study on transfer learning capabilities for pneumonia classification in chest-X-rays image, 2021, arXiv preprint arXiv:2110.02780.
[44] T. Zebin, S. Rezvy, COVID-19 detection and disease progression visualization: Deep learning on chest X-rays for classification and coarse localization, Appl. Intell. 51 (2) (2021) 1010–1021.
[45] L.O. Teixeira, R.M. Pereira, D. Bertolini, L.S. Oliveira, L. Nanni, G.D. Cavalcanti, Y.M. Costa, Impact of lung segmentation on the diagnosis and explanation of COVID-19 in chest X-ray images, Sensors 21 (21) (2021) 7116.
[46] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.
[47] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2097–2106.
[48] V. Teixeira, L. Braz, H. Pedrini, Z. Dias, Dualanet: Dual lesion attention network for thoracic disease classification in chest X-rays, in: 2020 International Conference on Systems, Signals and Image Processing, IWSSIP, IEEE, 2020, pp. 69–74.
[49] I. Allaouzi, M.B. Ahmed, A novel approach for multi-label chest X-ray classification of common thorax diseases, IEEE Access 7 (2019) 64279–64288.
[50] M.M.A. Monshi, J. Poon, V. Chung, F.M. Monshi, Labeling chest X-Ray reports using deep learning, in: International Conference on Artificial Neural Networks, Springer, 2021, pp. 684–694.
[51] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A.Y. Ng, M.P. Lungren, CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT, 2020, arXiv preprint arXiv:2004.09167.
[52] W. Boag, T.-M.H. Hsu, M. McDermott, G. Berner, E. Alesentzer, P. Szolovits, Baselines for chest X-ray report generation, in: Machine Learning for Health Workshop, PMLR, 2020, pp. 126–140.
[53] S. Jain, A. Smit, S.Q. Truong, C.D. Nguyen, M.-T. Huynh, M. Jain, V.A. Young, A.Y. Ng, M.P. Lungren, P. Rajpurkar, VisualCheXbert: Addressing the discrepancy between radiology report labels and image labels, in: Proceedings of the Conference on Health, Inference, and Learning, 2021, pp. 105–115.
[54] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, ChestX-ray: Hospital-scale chest X-ray database and benchmarks on weakly supervised classification and localization of common thorax diseases, in: Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics, Springer, 2019, p. 369.
[55] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, 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, no. 01, 2019, pp. 590–597.
[56] H. Wang, Y.-Y. Yang, Y. Pan, P. Han, Z.-X. Li, H.-G. Huang, S.-Z. Zhu, Detecting thoracic diseases via representation learning with adaptive sampling, Neurocomputing 406 (2020) 354–360.
[57] S. Albahli, H.T. Rauf, A. Algosaibi, V.E. Balas, AI-driven deep CNN approach for multi-label pathology classification using chest X-rays, PeerJ Comput. Sci. 7 (2021) e495.
[58] K. Almezhghwi, S. Serte, F. Al-Turjman, Convolutional neural networks for the classification of chest X-rays in the IoT era, Multimedia Tools Appl. 80 (19) (2021) 29051–29065.
[59] L. Seyyed-Kalantari, G. Liu, M. McDermott, I.Y. Chen, M. Ghassemi, CheXclusion: Fairness gaps in deep chest X-ray classifiers, in: BIOCOMPUTING 2021: Proceedings of the Pacific Symposium, World Scientific, 2020, pp. 232–243.
[60] J.P. Cohen, M. Hashir, R. Brooks, H. Bertrand, On the limits of cross-domain generalization in automated X-ray prediction, in: Medical Imaging with Deep Learning, PMLR, 2020, pp. 136–155.
[61] M. Hashir, H. Bertrand, J.P. Cohen, Quantifying the value of lateral views in deep learning for chest X-rays, in: Medical Imaging with Deep Learning, PMLR, 2020, pp. 288–303.
[62] E.H. Pooch, P. Ballester, R.C. Barros, Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification, in: International Workshop on Thoracic Image Analysis, Springer, 2020, pp. 74–83.
[63] J.P. Cohen, P. Morrison, L. Dao, K. Roth, T.Q. Duong, M. Ghassemi, COVID-19 image data collection: Prospective predictions are the future, 2020, arXiv preprint arXiv:2006.11988.
[64] Y. Wang, L. Sun, Q. Jin, Enhanced diagnosis of pneumothorax with an improved real-time augmentation for imbalanced chest X-rays data based on DCNN, IEEE/ACM Trans. Comput. Biol. Bioinform. 18 (3) (2019) 951–962.
[65] F. Charte, A.J. Rivera, M.J. del Jesus, F. Herrera, MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation, Knowl.-Based Syst. 89 (2015) 385–397.
[66] H. Salehinejad, E. Colak, T. Dowdell, J. Barfett, S. Valaee, Synthesizing chest X-ray pathology for training deep convolutional neural networks, IEEE Trans. Med. Imaging 38 (5) (2018) 1197–1206.
[67] W. Qu, I. Balki, M. Mendez, J. Valen, J. Levman, P.N. Tyrrell, Assessing and mitigating the effects of class imbalance in machine learning with application to X-ray imaging, Int. J. Comput. Assist. Radiol. Surg. 15 (12) (2020) 2041–2048.

[68] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al., Chexnet: Radiologist-level pneumonia detection on chest X-rays with deep learning, 2017, arXiv preprint arXiv:1711.05225.
[69] K.F. Monowar, M.A.M. Hasan, J. Shin, Lung opacity classification with Convolutional Neural Networks using chest X-rays, in: 2020 11th International Conference on Electrical and Computer Engineering, ICECE, IEEE, 2020, pp. 169–172.
[70] Z. Ge, D. Mahapatra, S. Sedai, R. Garnavi, R. Chakravorty, Chest X-rays classification: A multi-label and fine-grained problem, 2018, arXiv preprint arXiv:1807.07247.
[71] Z. Huang, D. Fu, Diagnose chest pathology in X-ray images by learning multi-attention Convolutional Neural Network, in: 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference, ITAIC, IEEE, 2019, pp. 294–299.
[72] K. Wang, X. Zhang, S. Huang, KGZNet: Knowledge-guided deep zoom neural networks for thoracic disease classification, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2019, pp. 1396–1401.
[73] R. Qin, K. Qiao, L. Wang, L. Zeng, J. Chen, B. Yan, Weighted focal loss: An effective loss function to overcome unbalance problem of chest X-ray14, IOP Conf. Ser.: Mater. Sci. Eng. 428 (1) (2018) 012022.
[74] J. Islam, Y. Zhang, Towards robust lung segmentation in chest radiographs with deep learning, 2018, arXiv preprint arXiv:1811.12638.
[75] S. Reza, O.B. Amin, M. Hashem, TransResUNet: Improving U-net architecture for robust lungs segmentation in chest X-rays, in: 2020 IEEE Region 10 Symposium, TENSYMP, IEEE, 2020, pp. 1592–1595.
[76] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.
[77] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[78] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[79] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.
[80] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.
[81] V.-L. Nguyen, E. Hüllermeier, M. Rapp, E. Loza Mencía, J. Fürnkranz, On aggregation in ensembles of multilabel classifiers, in: International Conference on Discovery Science, Springer, 2020, pp. 533–547.
[82] F. Chollet, et al., Keras, 2015, GitHub, URL: https://github.com/fchollet/keras.
[83] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, et al., Tensorflow: Large-scale machine learning on heterogeneous distributed systems, 2016, arXiv preprint arXiv:1603.04467.
[84] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[85] G. Bradski, The OpenCV Library, Dr. Dobb's J. Softw. Tools (2000).
[86] F. Charte, A.J. Rivera, M.J. del Jesus, F. Herrera, REMEDIAL-HwR: Tackling multilabel imbalance through label decoupling and data resampling hybridization, Neurocomputing 326 (2019) 110–122.
[87] G. Rimeika, E. Mockiene, et al., Deep learning model for chest X-ray pathology classification performance on an independent spanish dataset, in: European Congress of Radiology-ECR 2020, 2020.
[88] T. Zhang, X. Li, Z. Qu, Lesion attentive thoracic disease diagnosis with large decision margin loss, Biomed. Signal Process. Control 71 (2022) 103202.
[89] U. Kamal, M. Zunaed, N.B. Nizam, T. Hasan, Anatomy-xnet: An anatomy aware Convolutional Neural Network for thoracic disease classification in chest X-rays, IEEE J. Biomed. Health Inf. 26 (11) (2022) 5518–5528.
[90] H. Liz, M. Sánchez-Montañés, A. Tagarro, S. Domínguez-Rodríguez, R. Dagan, D. Camacho, Ensembles of Convolutional Neural Network models for pediatric pneumonia diagnosis, Future Gener. Comput. Syst. 122 (2021) 220–233.
[91] A.I. Aviles-Rivero, P. Sellars, C.-B. Schönlieb, N. Papadakis, GraphXCOVID: Explainable deep graph diffusion pseudo-labelling for identifying COVID-19 on chest X-rays, Pattern Recognit. 122 (2022) 108274.

Helena Liz is an Associate Researcher at Universidad Rey Juan Carlos (URJC) Computer Science Department under project CYNAMON. She did her undergraduate studies in Biology at Universidad Autónoma de Madrid and she has an M.Sc in Bioinformatics and Computational Biology at Universidad Autónoma de Madrid. Currently, she is Ph.D candidate at Universidad Politécnica de Madrid. Her research interests include Deep Learning, AI and Machine Learning applications in medicine, among others. Contact her at: helena.liz@urjc.es.

Dr. Javier Huertas-Tato obtained his PhD in Computer Science at Universidad Carlos III de Madrid under a FPI research grant. Currently, he is working as a Ph.D. assistant lecturer at Universidad Politecnica de Madrid and collaborating with national and international research projects such as CIVIC, FightDIS, and IBERIFIER. His current research topics are disinformation detection, tracking, and countering; machine learning applied to environmental issues; and deep learning techniques such as convolutional networks and transformers.

Manuel A. Sánchez-Montañés received his B.Sc. degree (with honors) in Physics from the Universidad Complutense de Madrid, Spain, 1997, and Ph.D. degree (cum laude) in Computer Science from the Universidad Autónoma de Madrid, Spain, 2003. He is currently working with the Computer Science Department, Universidad Autónoma de Madrid. His research activity is focused in Artificial Intelligence and Advanced Data Analysis, carrying out theoretical developments and applications.

Javier Del Ser received his first PhD in Telecommunication Engineering from the University of Navarra (2006), and a second PhD in Computational Intelligence from the University of Alcala (2013). Currently he is a Research Professor in Artificial Intelligence at TECNALIA (Spain) and an adjunct professor at the University of the Basque Country (UPV/EHU). His research interests gravitate on Artificial Intelligence for data modelling and optimisation tasks in a diverse range of application fields (Energy, Transport, Telecommunications, Health and Industry, etc.). He also serves as an associate editor in a number of indexed journals, including Information Fusion, Swarm and Evolutionary Computation and IEEE Transactions on Intelligent Transportation Systems. He is a IEEE Senior Member, and has been included in the list of the 2% most influential authors in Artificial Intelligence in 2021 and 2022 elaborated by Stanford University.

David Camacho is currently working as Full Professor with the Departamento de Sistemas Informáticos at Universidad Politécnica de Madrid (Spain) and leads the Applied Intelligence and Data Analysis group (AIDA). He received a PhD in Computer Science (2001) from Universidad Carlos III de Madrid. His research interests include Data Mining, Evolutionary Computation, Social Network Analysis, and Swarm Intelligence, among others. Contact him at: david.camacho@upm.es.