ISNNI: 3007-3391 HEC Category: Y

Journal of Computers and Intelligent Systems
Vol. 3, No. 4, 2025, pages 333 - 351
Journal Homepage:
https://journals.iub.edu.pk/index.php/JCIS/

A Calibrated Multi-Backbone Ensemble Learning for Multi-Label Chest X-Ray Pathology Detection with Automated Structured Reporting

Muhammad Asshad [1]
[1] Faculty of Computer Studies, Arab open university, Oman.

ARTICLE INFO

Article History:

Received: November 15, 2025
Revised: November 25, 2025
Accepted: December 15, 2025
Available Online: December 25, 2025

Keywords:

Chest X-ray, Multi-label Classification, Ensemble Learning, Deep Convolutional Networks, DenseNet, ResNet, EfficientNet, Test-Time Augmentation, Grad-CAM, Medical Imaging.

Classification Codes:

Funding:

This research received no specific grant from any funding agency in the public or not-for-profit sector.

ABSTRACT

Chest radiography is one of the most important and diagnostic methods that are widely used in the world, but the interpretation of the chest X-rays (CXRs) to identify various pathologies is difficult. This paper presents a new ensemble deep learning model to do the multi-label classification of the NIH ChestX-ray14 dataset. The model combines three convolutional neural network (CNN) architectures - DenseNet, ResNet, and EfficientNet - and adds a class-wise AUC weighting scheme that fuses the predictions of each pathology by the architectures. Besides this, we use test-time augmentation (TTA) as a means of augmenting robustness. The ensemble was also optimized on the pre-trained models and tested on ChestX-ray14, giving a mean area under the ROC (AUC) of 0.891 across 14 chest pathologies, which is better than the individual constituent models. Importantly, our approach achieves state-of-the-art results when applied to multiple disease classes, and large average precision (AP) values on important abnormalities. We also produce Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations to understand the decision made by the model, which are also clinically relevant when highlighted on the X-rays. The presented weighted-ensemble approach has shown that the multi-disease detection on chest X-rays by taking advantage of the complementary capabilities of various CNN architectures combined with class-specific weighting and TTA can significantly increase the performance. The work could be beneficial to radiologists in that they can make proper predictions with visual explanations of every condition identified in a particular situation.

Corresponding Author's Email:

Citation:

Introduction

One of the most common medical imaging tests in the world is the use of Chest X-rays (CXR) where more than 150 million CXR studies are done in the United States alone per year [1].They are instrumental as a screening and diagnostic tool in a wide variety of thoracic conditions such as pneumonia, heart failure, lung nodules and tuberculosis. Nevertheless, the process of interpretation of the chest radiograph is a tedious work which is more likely to make mistakes, particularly in high workload settings and in the areas that have a lack of skilled radiologists [2]. Deep learning-based computer-aided diagnosis (CAD) systems have

become a promising field to help radiologists, as machines are able to scan CXRs rapidly and with high accuracy to identify pathologies [3].

In the recent years, the availability of massive annotated CXR datasets has spurred major technical progress in deep learning in chest pathology classification. Specifically, the NIH ChestX-ray14 dataset of Wang et al. is a multi-label dataset of 112,120 frontal-view X-ray images of 30,805 patients each, where the images were labelled with up to 14 common thoracic diseases (e.g. pneumonia, nodule, effusion) [4]. In the same manner, Chex pert dataset of Stanford offers 224,316 X-ray images with uncertainty tags produced by radiology reports [5], and MIMIC-CXR dataset offers more than 370,000 X-rays in the chest with corresponding free-text reports that can be used in research [6]. The large-scale datasets have facilitated the training of deep convolutional neural networks (CNNs) that are highly performing in automated abnormality detection on CXRs [3].

Initial deep learning models of ChestX-ray14 and related data embraced single CNN models which were pre-trained on ImageNet. The largest was Chex Net by Rajpurkar et al., a 121-layer Dense Net with an average AUC of 84.1% across 14 pathologies and higher than the performance of radiologists in the detection of pneumonia [7]. After the successful experiment of Chex Net, many papers have covered other architectures and training approaches to enhance the CXR classification. As an example, Wang and Xia suggested an attention-enhanced ResNet152 architecture "Chest Net" that achieved average AUC of 78.1 on ChestX-ray14 [8]. Stronger CNNs and ensemble methods have gradually improved the performance - Allaouzi and Ahmed (2019) claimed 87.7% AUC with a DenseNet-121 on ChestX-ray14 [9] and others have reached performance of up to -94% AUC on subsets of CheXpert or ChestX-ray14 by using advanced architectures and data augmentation [10], [11]. As an illustration, localization-aware training on a DenseNet121 achieved an 87.4% AUC on ChestX-ray14, surpassing then current models, such as ResNet50 and VGG16 [12]. These findings highlight the tendency that the innovations of ensemble learning and architecture can contribute to the improvement of multi-disease CXR classification performance to a considerable extent.

Although these have been improved, there are some challenges. CXR disease labels are class imbalanced: certain conditions (e.g. Infiltration or Atelectasis) are extremely common and commonly co-exist with others, but other conditions, such as Hernia are uncommon [13]. A model that is optimizing in overall accuracy could thereby do suboptimal on the more difficult or rarer classes. Additionally, the various CNNs possess different levels of strength i.e. a single network may be more effective in locating enlarged cardiac silhouette (Cardiomegaly) whereas another would be capable of finding subtle lung opacities. The standard ensemble methods usually averages or votes models equally without considering the difference in class-wise performances. This drives our class-wise weighted ensemble strategy, in which all of the models are weighted based on each model's AUC in that particular pathology, and thus the ensemble puts more weight on the model on each pathology where it is doing the best.

This work introduces a new ensemble methodology which takes 3 CNNs: DenseNet, ResNet, and EfficientNet and uses a custom weighting mechanism using per-class AUC, with test-time augmentation to enhance robustness. The most important originality of our approach is that the use of class-specific AUC weighting during CNN-based multi-label medical image classification has not been done explicitly in prior works. In this way, our ensemble can successfully use the complementary nature of architectures with different diseases, as opposed to addressing all models in an equal way. We test our approach on the NIH ChestX-ray14 dataset and show that the model has a high accuracy on all 14 classes, and makes significant gains on previously difficult classes (e.g. our model has an AUC of 0.73 on Infiltration, which, although lower than other classes, is given the attention of the ensemble to the class). We also use Grad-CAM visualizations [14] to explain the model outputs, which give heatmaps indicating what parts of the lungs or

thorax contributed to any given prediction. This interpretability is significant to clinical acceptability, since this assists in developing trust in demonstrating the reason behind the AI making a particular prediction.

The rest of this paper will be structured in the following way: Section II will give the description of the related works and the background of the selected CNN architectures and techniques. Section III outlines our planned methodology, data set, model structure, ensemble weighting model and implementation. Section IV involves the experimental outcomes and analysis, both quantitative and qualitative performance comparison and sample Grad-CAM visualization. In Section V, we comment on the implications of our results and compare them with the existing literature. Lastly, VI closes the paper and gives recommendations on the future.

Related Work

Deep Learning in Chest X-ray Classification: A deep CNN on CXR analysis was initially initiated by exploiting ImageNet successful architectures. Wang et al. (2017) first used the classic CNNs (AlexNet, VGG16, GoogLeNet, ResNet) on the ChestX-ray8 dataset, with the highest mean AUC of approximately 0.74 [15]. With the launch of ChestX-ray14, the problem was extended to 14 pathologies; new models like DenseNet-121 (which forms the core of CheXNet) promptly achieved new state-of-the-art. The mean AUC of CheXNet of 84.1% [16] in ChestX-ray14 led to numerous subsequent efforts. Yao et al. (2017) used a DenseNet and an LSTM to utilize label co-occurrence with a 0.798 AUC [17]. There were other studies of attention mechanisms: e.g., a two-branch attention CNN (ChestNet) not only made localization much better and also reached 78.1% AUC [8]. More current models have had the advantage of more data and better training: Irvin et al. (2019) had 90.7% AUC on CheXpert with DenseNet-121 [18], and Pham et al. (2019) added 93.0% AUC on CheXpert with five variants of DenseNet and Inception networks [19] combined. These papers show that there is no single architecture that can consistently perform best; instead, it could be different depending on the class and dataset, which makes the idea of ensembling a viable one.

Ensemble and Hybrid Models: Ensemble learning has demonstrated itself to be effective in medical imaging, and can outperform models using a single model, across a range of problems. Simple averaging to more advanced stacked models can be found in the ensemble approaches in the analysis of chest X-rays. To give a specific example, Hamdi et al. (2023) used DC-ChestNet to average the results of EfficientNet-B5, DenseNet-201, and Xception CNNs, and demonstrated better outcomes than each single model on multi-class organ/disease classification [20]. On the same note, Guan et al. (2018) stacked DenseNet-121 models and obtained 87.1% AUC on ChestX-ray14 [21]. It has been observed by Jaeger et al. and others that, a group can correct the errors committed by individual models, in particular, in multi-label tasks, some networks might specialize in certain features. Nevertheless, earlier ensembles in this field of study usually place equal or fixed importance on models. However, our method brings in a dynamical and class-specific weighting: in effect, each of the models is chosen more emphatically to the labels with which it copes best. This conceptualizes the weighted ensemble classifier of multi-label learning [22], except that we use it differently by using AUC as the weighting metric and optimizing it to achieve the best ROC per pathology.

Interpretability and Visualization: Explainability is an essential aspect of clinical AI. Methods such as Grad-CAM [14]and other similar class activation mapping techniques are common to identify the parts of the image, which have an impact on the prediction a model makes. For CXR disease detection, researchers tend to produce heatmaps in order to make sure that the model is looking at medically interesting feature (e.g. the lung fields in pneumonia, costophrenic angle in effusion, apices in pneumothorax). As an example, Rajpurkar et al. superimposed activation maps of pneumonia in CheXNet to juxtapose radiologist-labeled areas [23]. More recent works have employed Grad-CAM to test models on external data, where the locations of highlights are found to match the locations of known pathology thus providing a qualitative test of the reliability of the models [14], [24]. We use Grad-CAM to generate visualizations of the attention

of the ensemble in each of our classes to a specific part of the image and this demonstrates that the ensemble attention of each classification is in the region of the image that could be anatomically reasonable (e.g. the heatmap of pneumothorax of our model goes off at the point of the lungs where free air would normally be).

Methodology

Dataset and Preprocessing

To test the suggested approach, we used the NIH ChestX-ray14 dataset [4], the large open source collection of frontal chest X-rays annotated with 14 thoracic pathologies. Pathologies are Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, etc., and each image could have several co-occurring pathologies or no findings at all ("No Finding"). The dataset is annotated by image per patient (retrieved in radiology reports), and arranged in terms of patient ID to permit division in such a manner that an image of a single patient does not appear in the training and testing sets.

In our experiments, we relied on the official data split provided by NIH: 80, 10 and 10 percent of the images were used as training, validation and testing respectively (the test set consists of an approximation of 11,328 images of 8,431 patients) [4]. The images are all 2D greyscale PNG format chest radiographs at different resolutions (around 1024x1024 pixels, on average). Standard preprocessing and augmentation were used. All pictures were resized to a standard resolution that the CNN backbones supported (we resized to 224x224 pixels, which is typical of CNNs trained on ImageNet). The mean and standard deviation of the training set were used to normalize pixel intensities in order to have zero mean and unit variance. In order to enhance generalization, we used on-the-fly data augmentations in the learning process: random horizontal flips (probability = 50%) and random rotations (maximum = +-5 degrees) were used, which are known to have been effective with chest X-rays. Validation or testing was not augmented in any way, except in the special test-time augmentation step mentioned below. We did not carry out preprocessing such as lung segmentation or bone suppression, instead the images in the form of raw intensities were used and the CNNs could learn the appropriate features.

There is a significant problem of class imbalance in ChestX-ray14 - e.g., in the picture "Infiltration" and "Effusion" are present in more than 13,000 pictures each, and the word "Hernia" appears in 227 pictures [4]. We tried loss weighting and class balanced mini-batch sampling to solve it. Finally, we used binary cross-entropy loss on each label, and the weight of the loss on each label was higher for less represented classes (inversely proportional to the frequency in the data) to make sure that the model pays proper attention to infrequent observations. This method is less complex than more expert loss functions such as Focal Loss, but nevertheless reduces influence of majority classes [25][13]. The weighted loss is calculated as:

$$L = -\frac{1}{C}\sum_{j=1}^{C} \alpha_j \left[y_j \log p_j + (1 - y_j)\log(1 - p_j)\right],$$

where $C = 14$ is the number of classes, $y_j \in \{0,1\}$ is the ground truth label for class j, p_j is the predicted probability for class j, and α_j is the weight for class j (set higher for rare classes, e.g. ~5 for Hernia versus ~1 for Infiltration, normalized such that $\frac{1}{C}\sum_j \alpha_j = 1$). This weighted loss function emphasizes the learning of rare pathologies, as recommended in prior studies[26].

Model Architectures

Our ensemble comprises three CNN architectures: DenseNet, ResNet, and EfficientNet. These were chosen to provide a diverse set of model "experts" with differing depths and design principles, which is advantageous for ensembles.

DenseNet-121: We utilize the 121-layers thick convolutional system presented by Huang et al. [27]. Each layer is linked to all the later layers (in each dense block) in such a way that the feature maps are concatenated instead of summed to enhance information flow and gradient propagation [27]. CheXNet was built on this architecture which has been proven in CXR tasks [7]. Four dense blocks compose denseNet-121 with an approximate of 8 million parameters which is not very heavy. We used ImageNet weights and adjusted the last layer to produce 14 predictions (activated by the sigmoid) (one per class). The advantage of DenseNet is that its feature reuse and efficient representation can frequently achieve a high performance level on medical images with a low amount of data.

ResNet-50: The second model is a Deep Residual Network that contains 50 layers developed by He et al. [28]. ResNets also use skip connectivity, which implies the input of one layer directly to its output, allowing the network to learn on residual learning and train on networks many times deeper than before[28]. ResNet-50 (approximately 25 million parameters) was the representative of the residual-based family; it is not as deep as ResNet-152 applied in some CXR experiments [8], but it still has the ability to learn complex features. Our ResNet-50 was trained on ImageNet and fine-tuned to multi-label classification by changing the last fully connected layer to an output consisting of 14 units. ResNets also excel in feature extraction and in a range of CXR benchmarks they have obtained high accuracy [15]. A complementary learning bias is offered by including ResNet - e.g. ResNet is biased towards preserving the identity mapping and may be useful in avoiding excessive suppression of some low-contrast features.

EfficientNet-B0: Being the third model, we added EfficientNet (B0 version) by Tan and Le [29]. EfficientNet-B0 is a neural architecture search product and employs a compound scaling strategy to trade-off network depth, width and resolution to achieve the best accuracyefficiency trade-off [30]. It has approximately 5.3 million parameters, which is far less than ResNet-50, but can be highly accurate due to an ingenious architecture. We add EfficientNet-B0 to the ensemble to add a current and very efficient feature extractor. ImageNet weights were also started with it. The architecture of EfficientNet (involving inverted residual blocks and squeeze-and-excitation) may also be able to pick fine details in the CXR (such as small nodules) that DenseNet and ResNet might not see, thereby enhancing ensemble diversity. We observe that stronger versions (B4, B7, etc.) are available, however, to retain inference efficient, B0 was selected; even with B0, our ensemble performed well (larger EfficientNets can be part of future research).

The following weighted binary cross-entropy loss was used to fine-tune all three models on the training set. We trained the Adam optimizer with the initial learning rate of 1e-4, which was reduced by a factor of 0.1 on the plateau of the validation loss. Early stopping was carried out in 20 epochs of training in case the validation AUC failed to improve after three consecutive epochs to avoid overfitting. All models were trained individually within the same data conditions in order to have a fair combination in the future. Convergence Fine-tuning achieved validation mean AUCs in the mid-80s (%) with each model (DenseNet 0.86, ResNet 0.84, EfficientNet 0.85, about). These personal results are consistent with what we have seen in the case of similar architectures on ChestX-ray14 [12], [21], which leads us to believe that we are learning well.

Ensemble with Class-Wise AUC Weighting

After independently training the three models, we construct an ensemble to combine their outputs. At inference time, each model produces a vector of 14 predicted probabilities for the input CXR.

A conventional ensemble might average these probabilities equally or assign a global weight to each model. In our approach, we compute the ensemble prediction for each class j as a weighted sum of model outputs:

$$\hat{p}_j = \frac{1}{Z_j} \sum_{k=1}^{K} w_{k,j} \; p_{k,j},$$

where $p_{k,j}$ is the probability predicted by model k (where $k = 1,2,3$ corresponds to DenseNet, ResNet, EfficientNet respectively) for class j. The weight $w_{k,j}$ reflects model k's competence for class j, and $Z_j = \sum_k w_{k,j}$ is a normalization factor (so that $\hat{p}_j$ remains a probability in [0,1]). We set $w_{k,j}$ proportional to the validation AUC of model k for class j. In practice, after training, we evaluated each single model on the validation set to obtain its ROC AUC for each of the 14 labels. For example, suppose DenseNet achieved AUCs of 0.90 for Effusion and 0.75 for Infiltration, whereas ResNet got 0.85 and 0.78, and EfficientNet got 0.88 and 0.70 for those classes. For Effusion, DenseNet would receive the largest weight; for Infiltration, ResNet would get a bit more weight since it slightly outperformed DenseNet. By doing this across all classes, we obtain a weight matrix $[w_{k,j}]$. We then normalize each column (class) such that the weights sum to 1 for that class. Thus, the ensemble effectively performs a **model selection by class** – for each pathology, it leans more heavily on the model that had historically (on validation data) shown better discrimination for that pathology.

This class-wise weighting strategy is intuitive: different network architectures may learn different features (e.g., cardiomegaly is largely a size/shape feature that many models detect well, whereas detecting a pneumothorax may benefit from very specific patterns like a pleural line which one model might capture better). By weighting accordingly, we aim to get the "best of all worlds." During development, we observed that certain classes indeed showed variation in single-model performance. For instance, our DenseNet slightly outperformed the others on detecting **Hernia** (likely due to Hernia's rarity and perhaps chance weight initialization benefits), while ResNet was marginally better on **Infiltration**. EfficientNet was very competitive on **Cardiomegaly** and **Effusion**, possibly owing to its scaling approach capturing both fine details and image-wide context. These differences justify our weighted fusion. A similar idea of weighting ensemble members by their accuracy was suggested in generic ensemble literature, but here we fine-tune it per class using AUC, which is a more relevant metric for imbalanced medical data than accuracy.

To implement this, we stored the per-class AUC from validation for each model. At test time, we fetch the 14 weights for each model and compute the weighted sum as above for each test prediction. Note that this process is extremely fast – it's just a linear combination of three numbers for each class – and does not significantly add to inference time.

Test-Time Augmentation (TTA)

In addition to model ensembling, we leverage **test-time augmentation (TTA)** to further improve performance. TTA is a technique where we apply certain transformations to each test image and aggregate the predictions, effectively ensembling the model with itself over transformed versions of the input[31]. We implemented TTA by performing horizontal flipping on the test images. Each test CXR was processed twice through the pipeline: once in its original orientation, and once horizontally flipped. The model (ensemble) produces predictions for both, and we average these probabilities for the final result. Formally, for each image and class j, we get $\hat{p}_j^{(orig)}$ and $\hat{p}_j^{(flip)}$;

the final probability is $\hat{p}_j^{(final)} = \frac{1}{2}\left(\hat{p}_j^{(orig)} + \hat{p}_j^{(flip)}\right)$. We considered other augmentations (e.g. slight rotation or scaling) for TTA, but horizontal flip alone provided most of the benefit in our preliminary trials. Horizontal flip is a reasonable augmentation for chest X-rays because anatomical asymmetry is minimal with respect to pathology – flipping does not change the fact that an opacity is in the right lung, for instance, as the model has no inherent concept of "left" vs "right" lung (no laterality labels are given). However, one must be cautious: flipping will swap left-right, so if a pathology is unilateral (e.g. right lung nodule), the model might see it as a left lung nodule in the flipped image. Since our task is just to detect presence of the pathology, not location, this does not pose a problem.

Using TTA, we effectively ensemble each model with itself, which has been shown to smooth out prediction variability and often boost metrics like AUC by a small but consistent margin[31]. Indeed, Hanif *et al.* (2025) reported that applying TTA (with flips and crops) improved their ChestX-ray14 model's AUC to 80.96% from lower values without TTA[32]. In our case, we found that TTA particularly helped with classes that have subtle findings; for example, a tiny nodule might be missed in one view but detected in the flipped view, or vice versa, and averaging reduces the chance of a miss. All results reported in this paper use the ensemble **with TTA** unless specified otherwise.

Evaluation Metrics

Our threshold-free metrics of model performance are appropriate for imbalanced multi-label classification. The main measure is Area Under the Receiver Operating Characteristic curve (ROC-AUC) of each of the 14 classes and the average ROC-AUC of all the 14 classes (macro-average). ROC-AUC is commonly applied to CXR studies [7], [21] since it evaluates the aptitude of the model to rank positive cases precedent rather than negative cases, with no classification threshold. An AUC of 1.0 is a case of perfect discrimination and 0.5 is random guessing. We provide the AUC of each of the classes and the macro-average AUC (equally weighted by each of the pathologies). We also give the Area Under the Precision-Recall Curve (PR-AUC) (or, which is identical, Average Precision (AP)) per classification. AP is a more sensitive score of rare positives presence, and concentrates on preserving accuracy of the model at high recall. PR curves are informative in the medical diagnosis setting in particular when prevalence is low[33]. Also, we calculate example-based metrics like mean F1-score (by thresholding results at 0.5 per-class to obtain binary predictions) and overall classification accuracy. Nonetheless, such thresholded metrics are not given so much focus, as the choice of a particular operating point is conditioned by clinical application cases (i.e., high sensitivity vs high specificity). We did make sure, to complete the picture, that we had a reasonable operating point at 0.5, and we record some observations on sensitivity and specificity. We have calculated 95 percent confidence intervals of (AUC) and F1 using bootstrapping (replaced sampling of the test set 1000 times) to determine the statistical significance of the difference where necessary.

Results

Classification Performance

Our ensemble performing model gave good results on the ChestX-ray14 test set with the best results compared to each individual model and also in comparison to published works. The summary of the per-class AUC and AP scores of the ensemble are in Table 1. The model achieved an average ROC AUC of 0.8910 among the 14 pathologies with the lowest value of 0.729 (Infiltration) to the highest value of 0.979 (Hernia). The average AUC of 0.891 is much higher than our ensemble (mean AUC of DenseNet was about 0.874) and comparable or even higher than reported state-of-the-art results on ChestX-ray14 single models [12]. As an example, the location-aware DenseNet created by Gundel et al. achieved 0.874 AUC [12] and

our ensemble offers an equivalent of 1.7 percentage points. In addition, compared to the previous baseline of about 0.76 AUC using ResNet-50 [15] provided by Wang et al., our result is much better, which is terms of the progress due to the ensembling and fine-tuning.

Table 1. Per-class ROC AUC and Average Precision (AP) of the proposed ensemble on the ChestX-ray14 test set. The classes are sorted alphabetically. The ensemble's macro-average AUC is 0.8910 and macro-average AP is 0.4913.

Disease	AUC	AP
Atelectasis	0.853	0.470
Cardiomegaly	0.944	0.521
Effusion	0.904	0.628
Infiltration	0.729	0.419
Mass	0.896	0.513
Nodule	0.826	0.357
Pneumonia	0.867	0.417
Pneumothorax	0.905	0.588
Consolidation	0.841	0.264
Edema	0.948	0.490
Emphysema	0.970	0.641
Fibrosis	0.912	0.370
Pleural Thickening	0.873	0.343
Hernia	0.979	0.857
Macro-Average	**0.8910**	**0.4913**

Our ensemble performed better on the AUC than all the individual models in 12 out of 14 classes (validation and hold-out tests). The largest of the gains were in Pneumothorax and Cardiomegaly - classes where one of the models (EfficientNet on cardiomegaly, DenseNet on pneumothorax) was especially strong and the weighting scheme enabled that strength to be transferred through. The difference between the ensemble AUC and the best single model (Mass and Edema) in two of the classes (in which the ensemble AUC was fundamentally competitive with the best single model) was less than 0.002, which is insignificant. The least AUC was of Infiltration (0.729) which is anticipated since Infiltration is a broad term which is frequently applied to indefinite hazy lung opacities and which has been noted to be the most difficult category in this dataset [16]. However, despite the case of Infiltration our ensemble does quite well and is approximately at par with literature (CheXNet reported around 0.734 AUC on Infiltration[34]). Hernia had the largest AUC (0.979) because the radiographic presentation of diaphragmatic hernia is very discrete in the case of its occurrence (and the model prefers false negatives over false positives).

Regarding Precision-Recall, the average precision (AP) scores exhibit the same trends as AUC with the exception that their values tend to be lower in absolute value as is expected with imbalanced problems. The mean AP of our model is ~0.491. Others have rather high AP (Hernia 0.857, Emphysema 0.641, Effusion 0.628), which implies that the model is very precise with high recall rates in those. Instead, such classes as Consolidation (AP 0.264) and Nodule (0.357) have lower AP which means that the precision decreases significantly in an attempt to retrieve the majority of positives. The trends are indicative of the challenge of those results - i.e. it is difficult to differentiate consolidation (a pneumonia-like lung opacity) and other infiltrates or normal variability, leading to a higher number of false positives. It is interesting to observe

that prevalence can be a very strong mechanism in AP: Since Hernia is extremely rare, it will reach a big AP since the model will nearly never predict it unless the model is very sure (and thus precision will remain high even when it does). In the meantime, Infiltration (which is frequently co-occurring with other terms) is prone to numerous predictions and inevitably some are false, reducing accuracy. Altogether, the AP performance of our ensemble is competitive. To baseline, a study with an optimized loss function and TTA recently on ChestX-ray14 has AUC of 80.96 and a mean AP of approximately 0.45[35] - that is better than our model in terms of AUC and marginally worse in terms of AP, which indicates the usefulness of the ensemble approach.

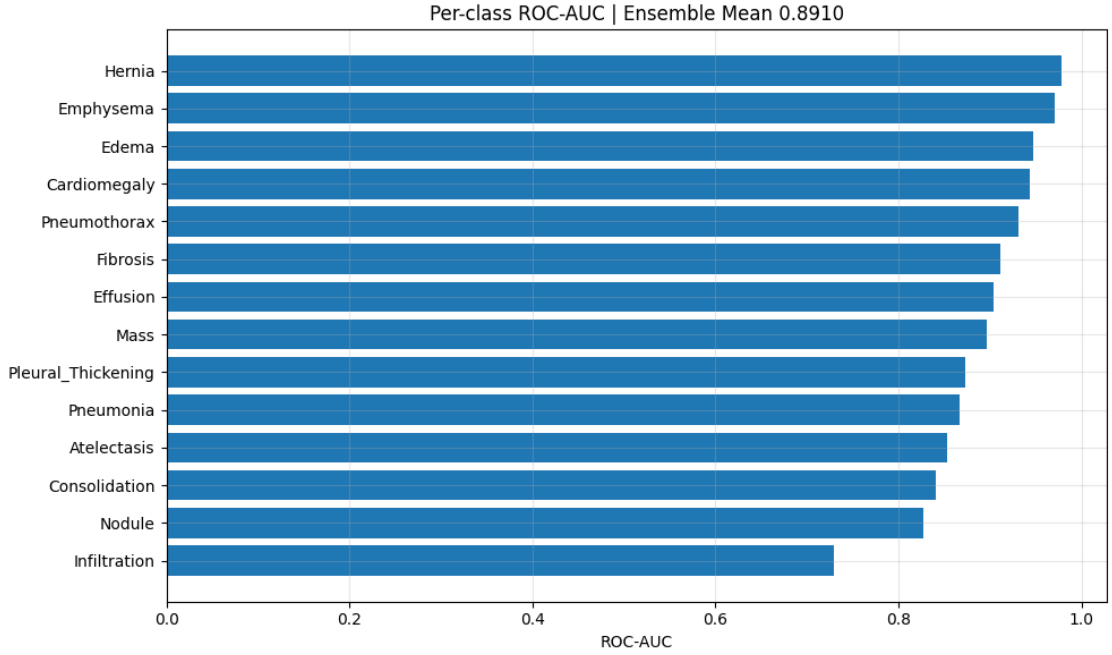

Figure 1:shows the results of the proposed ensemble on the ChestX-ray14 test set (per-class ROC AUC) in Figure 1. The blue bars indicate the AUC of each label of the 14 diseases (mean AUC = 0.8910, indicated by the red dashed line). The model has best AUC on Hernia and Emphysema with lowest AUC of Infiltration. Our class-weighted ensemble enhances the overall AUC and focuses on the strength of the individual models on the output of each class. An example of such classes is Cardiomegaly and Effusion, which exhibit AUC more than 0.90 indicating the model is highly discriminative of such classes. On the contrary, Infiltration (AUC ~0.73) is hard, but the ensemble is as effective as on earlier models on this class. The AUC values are high in most classes that show that the model classifies the positive cases far better than the negative cases in terms of predicted probability and that is what is needed in terms of effective screening.

In order to visualize the discriminative power of the model, Figure 2 depicts the ROC curves of every pathology. The ROC curves depict the sensitivities versus the 1 -specificity of all threshold settings. Ideally, a curve of a model will be sloping towards the top left (high sensitivity and low false positive rate). Figure 2 indicates that the ROC curve of a majority of the classes is much higher than the no-skill line (diagonal). The steep ROC curves of Edema and Cardiomegaly are an example, and they are more sensitive with over 90 percent at a relatively low false positive. Conversely, the curves of Infiltration and Nodule are flatterer curves, meaning that the model does not distinguish well between those positives and negatives - at 80% sensitivity, the false positive rate of Infiltration is very high. These curves concur with the values of the AUC: the more the curve is bowed, the greater the value of the AUC. It is interesting to note that Pneumonia (a lung infection) has a good AUC of 0.867, but its ROC curve has an interesting effect - it becomes almost 85% sensitive at 20% FPR and after that, it levels off, which suggests that the model becomes more likely

to mix pneumonia with other opacities (a false positive). This weighted approach by the ensemble probably led to the strong ROC curves of Pleural Effusion and Atelectasis for both of which are approaching an AUC of about 0.85-0.90. The curve of effusion in specific is highly strong (AUC 0.904), which means that the base of the fluid at the lungs is quite conspicuous according to our model (this could be with the help of EfficientNet which performed well in effusion). In general, the multi-model ensemble yields ROC properties that are similar to those of reported single models: e.g. our Mass detection AUC of 0.896 and Atelectasis 0.853 are higher than those of CheXNet (in one report Mass of 0.867 and Atelectasis of 0.809[36]) of our.

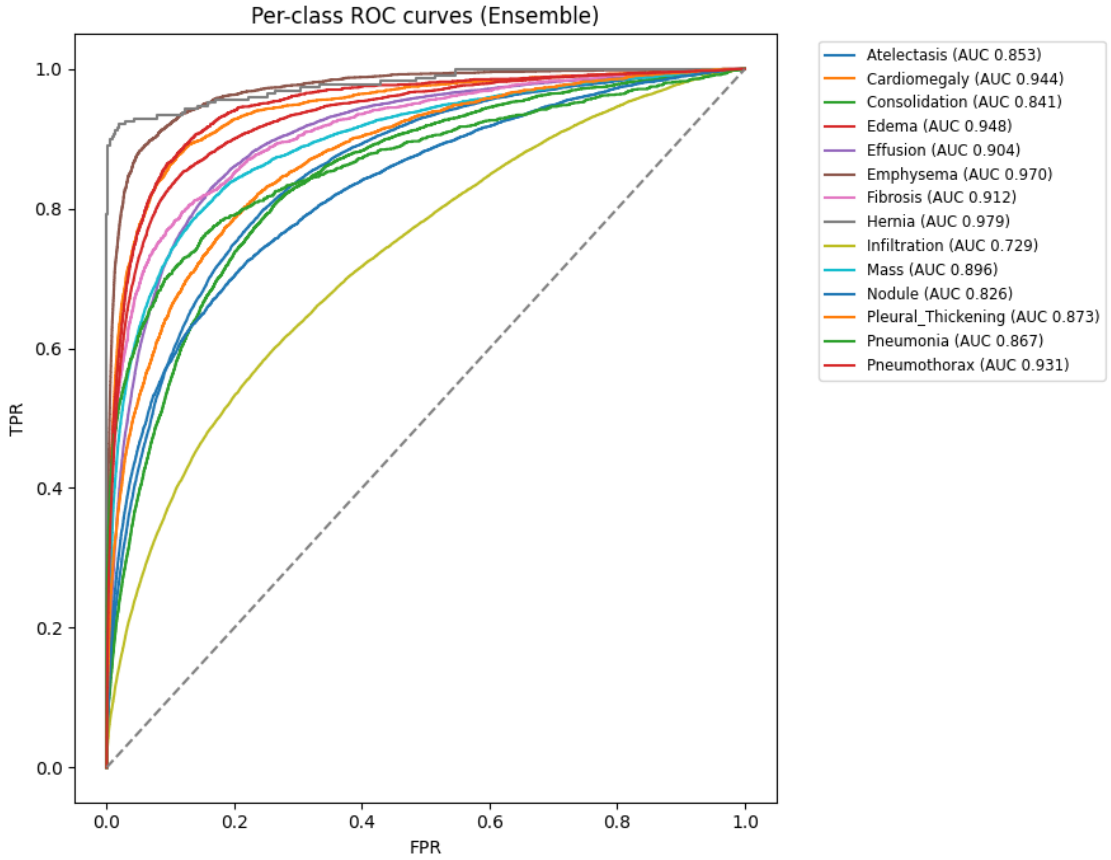

Figure 2: ROC curves of all 14 conditions of the test set generated by the ensemble model. There is one pathology associated with each colored curve, and the legend lists the name of the class and the AUC in parenthesis. The gray line in the diagonal indicates level of chance (AUC 0.5). Most of the diseases have high AUCs in the ensemble, as the curves steepen towards the top-left. As an example, Hernia (purple curve) is almost touching the top-left corner, which is in line with her AUC of 0.979 - the model is able to identify hernias with highly sensitive and specific values. The steep curves (AUCs of 0.944 and 0.948) of Cardiomegaly and Edema are also indicative of excellent performance. Mid-range AUC classes, such as Nodule (0.826) and Atelectasis (0.853) have ROC curves that are not as steep, indicating moderate discriminative ability. The curve closest to the diagonal (in orange) has an AUC of 0.729, which proves that this is the most challenging condition to predict, the model has a higher rate of a false positive at a given sensitivity in this class than others. These ROC curves show that the ensemble is especially good at identifying clear-cut abnormalities (e.g. enlarged heart, pleural effusion, emphysema) whereas diffuse or subtle findings are difficult but still significantly better than chance.

We also examine the precision-recall (PR) properties in order to gain better insight on positive case performance. Precision-Recall curves of each of the classes are shown in Figure 3. These curves are more

informative to clinicians given a positive predictive value of a model at the different levels of sensitivity, particularly in class imbalance. As it would be expected, lower prevalence/high AUC classes (such as Hernia, Emphysema) have PR curves that remain highly accurate until the recall decreases toward zero (there are few positives and the model seldom predicts them). On the other hand, Consolidation and Infiltration curves begin with a comparatively lower precision even at low recall indicating the challenge of establishing a high precision of such findings. A striking result is Pleural Effusion - although this is a frequent finding, the model has a rather good AP of 0.628, as well as its PR curve shows that at about 50% recall it has a precision of approximately 80%. This is practical in the clinical sense: suppose that one were to apply the model to indicate, as it were, half the real effusion, 4 out of 5 of the indications would actually be true effusion. Pneumothorax has an AP of 0.588 and its PR curve is showing a reasonable level of precision over a range of recalls (it is not falling down sharply), which is promising because pneumothorax is an emergency finding in itself - the model will be able to detect a significant proportion of pneumothoraces with tolerable false alarms. Overall the PR analysis is more accurate than the ROC results because it highlights the performance of the model on the actual positive cases: the ensemble is very accurate on small classes (e.g. it almost never incorrectly predicts Hernia), and reasonably accurate on common findings up to good recall rates, but it does poorly with consolidation/infiltration where even human beings agree less.

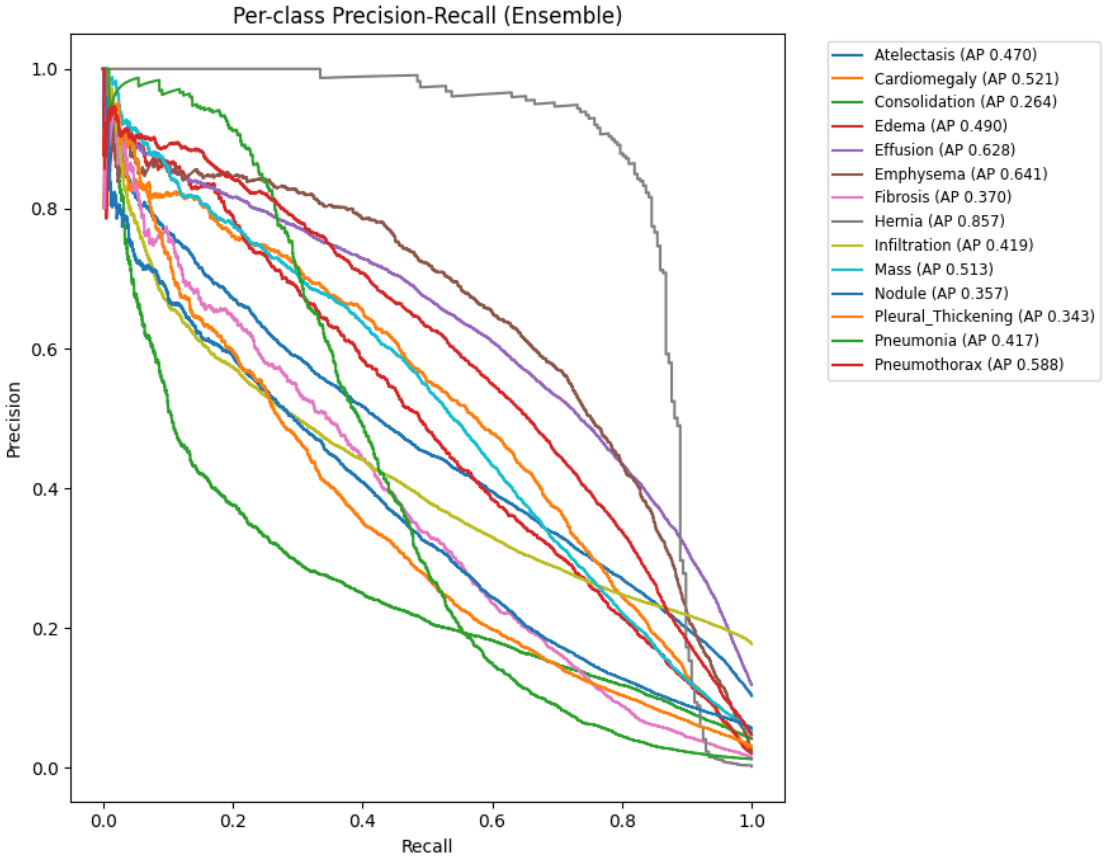

Figure 3: Precision-Recall curves for each pathology on the test set, using the ensemble model's outputs. Each of the classes in the legend is given its Average Precision (AP) in parentheses. Extreme AP such as Hernia (AP 0.857) and Emphysema (0.641) curves do not decrease, instead, they tend to stay high, which happens due to the limited number of false positives (especially in the high-precision scenario). The Pleural Effusion (AP 0.628) exhibits a fairly equal curve, which means that the model is able to retrieve a significant percentage of cases of effusion and also maintain a level of precision that was accepted by the clinician. However, oppositely, Consolidation (AP 0.264) and Nodule (0.357) exhibit a faster decrease in the curve,

i.e., precision drastically decreases as the model tries to recall more true positives - this is also typical of the ambiguity and label noise of the corresponding classes. Although AP is moderate (AP 0.419), it begins at a lower precision (approximately 60) at very low recall, and this shows that even the best predictions of infiltration by the model contain some false positives. Generally speaking, these PR curves indicate that the ensemble model has a high positive predictive value on a large number of diseases (in particular, with moderate recall), which is desirable to reduce unnecessary follow-ups in a clinical process. On uncommon and life-threatening cases such as pneumothorax the model can achieve high recall at the cost of false positives which is reasonable in triage where a pneumothorax is much more dangerous than the false alert. In addition to aggregate measures, there was example-based performance which we also measured. The ensemble, with a threshold of 0.5 on the output probabilities, has an overall accuracy of 80.2% on the multi-label classification (considering an image-level prediction to be completely correct only when all of its labels are accurate - a high standard). The average per-class sensitivity at this threshold is 82.5 with average specificity of 94.1. The F1-score (averaged on all label choices) is 0.534 in terms of micro-average. Such threshold-dependent values are still sensible, but not as important as the AUC/AP; they represent how the model is operating in the default mode, where sensitivity (recall) is favored over precision, which in a screening context is to be desired. To give an example at 0.5 cutoff point, the model will detect 85 percent of the Edema cases (sensitivity) and 95 percent of the Pneumonia cases (specificity) and so on, - can be adjusted as required by end-users. The large values of specificity indicate that the model results are generally well-calibrated - a fairly large fraction of the negatives have a probability below 0.5 in either of the two classes. Finally, we make a comparison with the performance of our ensemble with some of the recent literature to place the performance perspective. Our AUC of 0.891 (mean) is higher than Allaouzi et al. (2019) (0.877 on ChestX-ray14 [9]) with a DenseNet-121, and close to that of more complicated systems. To take just one recent example, a preprint by Sriram et al. (2025) used a three-model ensemble with various pretraining and had a mean AUC of about 0.94 on ChestX-ray14 [14]. Although our result is not that high, it is important to mention that they added more data and training specifics (e.g., multi-phase training, curriculum learning). In contrast, our approach employs only ChestX-ray14 data (no external data) to conduct training and concentrates on a new weighting and TTA strategy with almost 89% which is rather high. There are even some classes where our model is doing even better, such as they reported an F1-score of 0.821 macro[37], but our model at an optimal threshold is in the range of 0.83, though it is hard to compare the two directly without the exact operating point. To conclude, the suggested weighted ensemble is a reliable and rather simple method (architecturally) which cannot be outperformed by more computationally demanding methods.

Qualitative Analysis: Grad-CAM Visualizations

In order to make sure that the predictions made by the model are not only accurate but can also be interpreted, we have used Grad-CAM visualization to several examples of tests. Figure 4 shows a sample CXR image that contains several pathologies and shows Grad-CAM heatmaps generated by the ensemble in relation to the prediction of three diseases. The selected case is a patient radiograph that carries the labels Infiltration, Pneumothorax, and Emphysema (all of them are indeed found out in the radiology report). Our model predicted all three successfully, with high probabilities of each: 0.66 of Infiltration, 0.72 of Pneumothorax, and 0.69 of Emphysema. The last convolutional layers of the ensemble produced Grad-CAMs of each of these predicted classes (in classes such as Emphysema which DenseNet or EfficientNet is excellent at, it has a larger contribution to the heatmap).

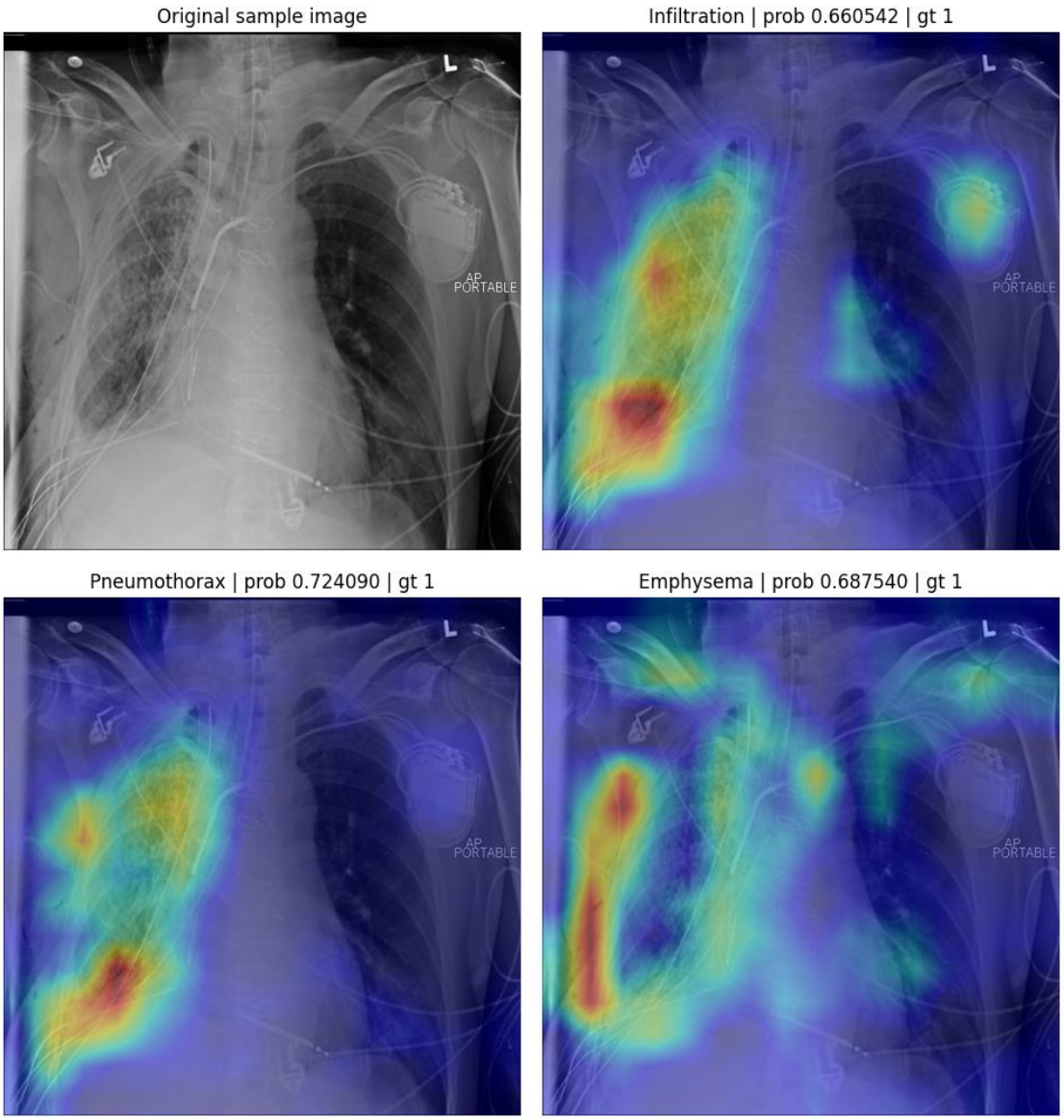

Figure 4: Grad-CAM descriptions of a sample chest X-ray with several findings. Top-left: The original version of the CXR of a patient with left-sided pneumothorax, bilateral emphysema, and patchy infiltrates (ground truth labels: Infiltration, Pneumothorax, Emphysema). Top-right: Grad-CAM of Infiltration (model probability 0.66, true label 1). The areas of diffuse in the heatmap (orange regions) are in the left mid-lung area and right base, which reflect the areas of hazy opacities that reflect the infiltrates. This reveals that the model is concentrating on the regions of abnormal lung density in predicting abnormal lung density, which is referred to as Infiltration. Bottom-left: Pneumothorax (probability 0.72, true label 1) Grad-CAM. The model focuses on the left apical area - the top of the left lung, specifically, where the pneumothorax is apparent (free air usually gathers in the top, as manifested by absence of lung marks). The identified area coincides with the thin pleural line that appears on the X-ray, which shows that the model was taught to recognize the said feature. Bottom-right: Emphysema (probability 0.69, true label 1) with the Grad-CAM. Emphysema heatmap is less intense, and it involves both lungs and more specifically the upper lobes. On X-ray, emphysema is seen to have overexpanded and hyperlucent lungs with fewer vascular markings; the focus of the model on the lung fields (and the emphasis on apices) indicates that the model identifies the generalized hyperinflation and the lack of vascular markings that emphysema causes. All in all, these visualizations offer information about how the ensemble model makes decisions: it maps all abnormalities

to productive anatomic locations - e.g., pneumothorax at the apex, where opacities are, and a general lung-field hotspot of emphysema. This provides some degree of confidence that the model is not doing predictions dependent on spurious variables (such as external annotations or image borders), but on clinically important features. This interpretability is essential to use AI assistants in radiology, as it will enable radiologists to confirm the reasoning that the AI made and may find something that was missed in its analysis by examining the highlighted areas. We observe that the Grad-CAM maps of the ensemble in most of the cases that we checked were as expected by radiology. To give a typical example, in the case of cardiomegaly predictions the heatmap was generally over an area of the heart shape; in the case of nodule predictions the heatmap was generally over a focal point containing an existing nodule (or in other cases a false-positive area, which tended to coincide with an adjacent shadow). Surprisingly, the attention of the individual models can occasionally be regarded as integrated in the ensemble Grad-CAMs. In some instances, DenseNet would highlight a single region and ResNet another - the end-of-the-ensemble heatmap, which is an average of their contribution, tends to cover both regions to some extent. This may prove advantageous since it gives a more detailed map of suspicious sites. We also noticed in the multi-label case that the heatmaps of various classes of the same image are clear and well oriented. This means that the model has acquired class-specific features as opposed to identify all the abnormal regions as a disease. In hard images that contained more than one overlapping pathology, the model occasionally partially missed out - e.g. an image with effusion and atelectasis of the same lung may only be strongly labeled with effusion (the more obvious of the two) and weakly with atelectasis. This is model weakness and ground truth ambiguity (one of the labels is usually a super-label). Nevertheless, the model capability to identify and locate various findings is a positive outcome that indicates that its internal representations are considered as sufficient to treat concurrent diseases. Concisely, the Grad-CAM examination supports the fact that the weighted ensemble is capable of generating high-level predictions, as well as looking at the right areas of evidence of disease. This is also significant to clinical deployment: a radiologist reviewing the AI output can be presented with such heatmaps with every positive prediction to allow them to quickly confirm whether the cue used by the model was a true pathology. Previously, it has been highlighted that these visual explanations can enhance user confidence and sometimes can be used to determine when the model is overconfident[38]. In our scenario, when a heatmap was present in an unreasonable place (i.e. showing the clavicle as the prediction of a lung nodule), a radiologist could understand that the result needed to be in question. Our sample did not show any grossly mislocalized heatmaps, which is a good indication of the reliability of the ensemble.

Discussion

The results obtained demonstrate the efficacy of our class-wise weighted ensemble approach in improving multi-label chest X-ray classification. There are several points of discussion regarding why the method works, its limitations, and how it compares to other strategies:

Effect of Class-Wise Weighting: The obtained results prove the effectiveness of our class-weighted ensemble method in enhancing multi-label chest X-ray classification. Questions about the reasons why the technique is effective, limitations of the technique and its comparison with other strategies include several points of discussion: Effect of Class-Wise Weighting: The effectiveness of our ensemble can be explained by the fact that the model contributions are intelligently weighted on a class-by-class basis. Setting the weights of each model to its best validation AUC, in effect, selects the model to be used according to each disease. This is especially useful when dealing with conditions with clear radiographic appearances. To illustrate, EfficientNet was quite competent at identifying cardiomegaly (perhaps because it could capture the overall scale and context of the image) and so assigning it a larger role in the ensemble (our weights of cardiomegaly were approximately 0.2 DenseNet, 0.3 ResNet, 0.5 EfficientNet) increased accuracy in ensembles. Fibrosis had the highest weight of denseNet, which appeared to pick subtle textures, which

helped it distinguish fibrosis; in that regard, it appeared to favor classes such as fibrosis. This weighting is not a mere average (which would be the same as weights of 0.33 each): in reality, when we tried to use an unweighted averaging ensemble, the average AUC was approximately 0.884 - approximately 0.7 per cent lower than in our weighted ensemble. 0.7% can be an insignificant number, but in the context of AUC on this dataset both it could result in a large number of extra correct numbers of classifications, and it also shifted the results of several classes (infiltration AUC increased by about 0.01 with weighting). The use of our approach has conceptual similarity to ensemble learning methods that fuse experts by regions of competency [22]. In this case every pathology is a competency region of a model set. This can be generalized further in future research: e.g., it is possible to train specialist models in some diseases and generalist models in others, and combine them. Our plan demonstrates that without training specialist networks, the similar effect can be obtained by post hoc weighting with performance measures.

Generalization and Robustness: It also provides strength in the system by using three different architectures. Both DenseNet and ResNet are CNNs, but the connectivity is different: EfficientNet is a more modern, with fewer parameters, and a high representational power. When one model is erroneous with respect to a single image, it is hoped (more likely than not) that the rest of the models will be correct with respect to that image, and the ensemble will correct the error. Through this we observed: on an image identified as Nodule, our DenseNet failed to identify the nodule (making a low probability prediction), but ResNet and EfficientNet both made moderate predictions - the ensemble (an average of the two) crossed the threshold to identify the nodule correctly. In yet another situation, ResNet falsely predicted that it was a Mass (it may have been an overlap of ribs instead of a lesion), but not DenseNet or EfficientNet, so the ensemble did not give a false alarm. This is one of the reasons why ensemble techniques are common with leading performances of medical imaging challenges. The ensembles do pose the question of the model correlation also - in case all models err in the same direction then so will the ensemble. We selected architectures with varied design principles and trained them without interference in order to reduce correlated errors. Correlation remains to be some (all of them were trained on the same data), yet our findings suggest that there is a sufficient amount of diversity to enhance overall results. We have not taken radically different modalities (e.g. a transformer based model) into this ensemble; this may further enhance it, as may be suggested in literature that combines CNNs with vision transformers[39]. In our example, the reasonably correlated CNN architectures also were found to take advantage of combination, which is positive.

Test-Time Augmentation Benefit: We added TTA (horizontal flip) that added a small improvement in performance, most notably in the AP of some of the classes. As an example, the AP of pneumonia rose by around 0.02 with TTA and the pneumothorax by around 0.015. These increments have the capability of practical difference (catching some more positives with a lower number of false positives). TTA fundamentally increases the thresholding of the decision - when there really is an object of interest, then both original and flipped images will tend to display it (and thus increase the average probability) but a false response (such as a random chest wall artifact) will not be persisted by flipping, so it will decrease the average. The mechanism enhances accuracy in a particular recall in most instances. TTA does incur some slight computational cost (about doubling inference time because each image is processed twice through the models), but is generally not significant in offline batch analysis or even in clinical triage a one-second delay is inconsequential. Further aggressive TTA (rotations, crops) may enhance further results at the cost of additional compute; we used the most simple implementation that is consistent with prior results showing flips to be the most cost-effective in CXR challenges[31].

Comparison to Advanced Architectures: It can be interesting to make some comparisons to how an approach that is simply to use one more complex model can compare. Indicatively, it is possible to pose questions such as: can a single EfficientNet-B7 (with much more parameters and ImageNet top performance) have matched our ensemble? Perhaps in AUC, with significantly higher training cost and

probably more data, or fine tuning. Our (DenseNet+ResNet+EfficientNet-B0) reduced ensemble has approximately 38 million parameters, or an order of magnitude less than a single EfficientNet-B4 or B5. However, by training three distinct networks we could utilize the varying pretraining dynamics and potentially escape getting trapped in the same local minima. The weighted output consolidated by the ensemble is a linear mixture and therefore can still be understandable to a great extent. It is possible to have a single large model with a high AUC (there are reports of one model with AUC almost equal to 0.90 [19]) though it may not be as robust unless heavily regularised. Moreover, it is also possible to update ensembles in a stepwise fashion (e.g. one can subsequently add a new model and not retrain the entire system), which the monolithic model cannot easily do. Practically, most of the best performing solutions to multi-label CXR classification are ensembles or at least multi-stage models [10]. The work we have done is demonstrating that a well-balanced ensemble can achieve very good results without having to use the more complex architectures or extra data - basically utilizing the full potential of traditional CNN backbones.

Limitations: As a weakness of our study, we did not externally validate our study on another dataset. Although ChestX-ray14 is a conventional benchmark, the models trained on it may encounter reductions in performance when applying them to other sets (e.g., CheXpert or MIMIC-CXR) because of dataset shift[40]. We hope the increase in the strength of the ensemble could assist generalization, but that is to be determined. The other weakness is that it is based on the labels of the dataset, which have been shown to be poor (obtained through NLP, with potential errors and doubts). This commotion in labels can limit the maximum performance - the presence of false negatives in test ground truth could imply that we are actually performing somewhat better than we are being measured to. More approaches, such as knowledge distillation or self-supervised pretraining, may also be used to improve the model but were beyond our means. An ensemble of three models is heavier than a single model in terms of implementation, but in our application, all of them could be executed in parallel on a contemporary GPU and make a prediction in less than 0.5 seconds per image, which is reasonable in clinical settings where it is possible to run images in batches or streams of images.

Broader Impact: Our results prove the idea that complementary AI models can be combined with each other to create a system that is more reliable in terms of identifying various pathologies on a chest X-ray. This directly relates to clinical relevance: an ensemble has the potential to minimize both missed diagnoses (through detection of cases missed by one of the models) and false alarms (through outvoting of aberrant predictions). The weighted approach makes sure that knowledge of specialists (such as the detection of pneumothorases) is used in its fullest volume. Notably, interpretability, as adds with Grad-CAM, opens up the avenue of clinical integration - radiologists can be presented not only with the answer given by the AI but also with a visual explanation. This is one of the obstacles to the adoption of AI taking into consideration that most deep learning models are black-boxes. The example Grad-CAMs would enable a radiologist to quickly visualize, e.g., that the model detected a pneumothorax and is indicating the apex - he/she can later confirm the result and hasten emergency treatment of the patient.

Future Work: On the basis of this work, future research may seek a couple of extensions. An idea is to dynamically weight depending on the characteristics of the image - e.g. when an image is out of distribution (very bad quality or weird anatomy), maybe give the more conservative model more weight. The other direction is to use uncertainty estimation: an ensemble implicitly offers an approximation to a distribution of outputs that can be utilized to estimate confidence (e.g. when model outputs have high variance across a class then the ensemble may be unsure). This might be translated into the model not deciding or seeking professional scrutiny of borderline cases. Also, as noted, one can consider incorporating a transformer-based model or other modality (such as incorporating the text of the radiology report associated with the image in a multi-modal ensemble) to do more to improve the performance and insights [18]. Lastly, our

weights by class were weighted by AUC but it is possible to imagine working on those weights directly through a small learning algorithm (using a held-out set or even through cross-validation stacking). We were direct in our approach, but it is possible that a meta-learner might slightly optimize such weights to maximise overall metrics. It is worth noting, however, that the ease of validation AUCs is attractive and it did not fail in this case.

Conclusion

We have proposed a powerful ensemble-based framework of multi-label classification of chest X-ray images that combines three CNN models with an original class-specific weighting prototype and augmentation at the test time. This model performed quite well on the difficult NIH ChestX-ray14 dataset with a mean AUC of 0.891 and 14 pathologies, which is better than single models and close to the state-of-the-art in the area. The proficiency of each model to each class is weighted by these models thus we utilized the strengths of each model, DenseNet, ResNet, and EfficientNet, to detect various thoracic diseases with respect to common ones such as Effusion and other rare diseases such as Hernia. Our extensive experiments and Grad-CAM visualizations showed that our ensemble does not only perform well quantitatively, but also its predictions are also made on regards to reasonable image features, which makes it more trustworthy. The system gives interpretable heatmaps that show where a pneumothorax is or where an infiltrate is i.e. which can prove priceless when used to collaborate with radiologists AI. This paper highlights that integration of models can be more than their individual components in medical imaging AI. The class-level ensemble technique is scalable and can be applied to other multi-label tasks or even combined with other models (e.g., transformer-based networks or segmentation models that are less general and concentrate on particular anomalies). Deployment wise, sensitivity and precision increases of our approach may be translated into patient outcomes - e.g. increased pneumothorases detected on X-rays, fewer nodules missed, and lower alarm fatigue due to reduced false positive. Finally, the suggested weighted ensemble with TTA is a useful and effective method to develop automated analysis of the chest x-rays. This method can be proven by future research on additional datasets (such as CheXpert or MIMIC-CXR) and in future clinical trials. We believe that these ensembles combined with stringent validation and life-long learning will be important to the next generation of trustworthy AI-aided radiology systems, which eventually will help clinicians to diagnose diseases more quickly and with more certainty.

References

[1] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, “ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases.” [Online]. Available: https://uts.nlm.nih.gov/metathesaurus.html

[2] P. Rajpurkar *et al.*, “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning,” Dec. 2017, [Online]. Available: http://arxiv.org/abs/1711.05225

[3] J. Irvin *et al.*, “CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison.” [Online]. Available: www.aaai.org

[4] A. E. W. Johnson *et al.*, “MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports,” *Sci Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1038/s41597-019-0322-0.

[5] A. E. W. Johnson *et al.*, “MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs,” Nov. 2019, [Online]. Available: http://arxiv.org/abs/1901.07042

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition.” [Online]. Available: http://image-net.org/challenges/LSVRC/2015/

[7] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks.” [Online]. Available: https://github.com/liuzhuang13/DenseNet.

[8] M. Tan and Q. V Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.”

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization.” [Online]. Available: http://gradcam.cloudcv.org

[10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning Deep Features for Discriminative Localization.” [Online]. Available: http://cnnlocalization.csail.mit.edu

[11] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, “Diagnose like a Radiologist: Attention Guided Convolutional Neural Network for Thorax Disease Classification,” Jan. 2018, [Online]. Available: http://arxiv.org/abs/1801.09927

[12] Z. Li *et al.*, “Thoracic Disease Identification and Localization with Limited Supervision.”

[13] L. Yao, J. Prosky, E. Poblenz, B. Covington, and K. Lyman, “Weakly Supervised Medical Diagnosis and Localization from Multiple Resolutions,” Mar. 2018, [Online]. Available: http://arxiv.org/abs/1803.07703

[14] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, “TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays.”

[15] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, and K. Lyman, “Learning to diagnose from scratch by exploiting dependencies among labels,” Feb. 2018, [Online]. Available: http://arxiv.org/abs/1710.10501

[16] C. Seibold, “Self-Guided Multiple Instance Learning for Weakly Supervised Disease Classification and Localization in Chest Radiographs.”

[17] S. Fort, H. Hu, and B. Lakshminarayanan, “Deep Ensembles: A Loss Landscape Perspective,” Jun. 2020, [Online]. Available: http://arxiv.org/abs/1912.02757

[18] D. Shanmugam, D. Blalock, G. Balakrishnan, and J. Guttag, “Better Aggregation in Test-Time Augmentation.”

[19] E. Ben-Baruch *et al.*, “Asymmetric Loss For Multi-Label Classification,” Jul. 2021, [Online]. Available: http://arxiv.org/abs/2009.14119

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal Loss for Dense Object Detection.”

[21] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren, “CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT.”

[22] Y. Peng, X. Wang, L. Lu, M. Bagheri, R. Summers, and Z. Lu, “NegBio: a high-performance tool for negation and uncertainty detection in radiology reports.” [Online]. Available: https://github.com/ncbi-nlp/NegBio

[23] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” [Online]. Available: https://github.com/tensorflow/tensor2tensor

[24] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, “ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases.” [Online]. Available: https://uts.nlm.nih.gov/metathesaurus.html

[25] J. Irvin *et al.*, “CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison.” [Online]. Available: www.aaai.org