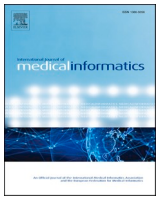
International Journal of Medical Informatics 178 (2023) 105211

Contents lists available at ScienceDirect

International Journal of Medical Informatics

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



Enabling chronic obstructive pulmonary disease diagnosis through chest X-rays: A multi-site and multi-modality study

Ryan Wang [a], Li-Ching Chen [a], Lama Moukheiber [b], Kenneth P. Seastedt [c], Mira Moukheiber [k], Dana Moukheiber [b], Zachary Zaiman [d], Sulaiman Moukheiber [e], Tess Litchman [f], Hari Trivedi [g], Rebecca Steinberg [h], Judy W. Gichoya [g], Po-Chih Kuo [a,*], Leo A. Celi [b,i,j]

[a] Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
[b] Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA
[c] Department of Surgery, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA
[d] Department of Computer Science, Emory University, Atlanta, GA, USA
[e] Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA, USA
[f] Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA
[g] Department of Radiology, Emory University, Atlanta, GA, USA
[h] Department of Medicine, Emory University, Atlanta, GA, USA
[i] Division of Pulmonary Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA
[j] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[k] The Picower Institute for Learning and Memory, Massachusetts Institute of Technology, Cambridge, MA, USA

ARTICLE INFO

Keywords:
Chronic obstructive pulmonary disease
Chest X-ray
Convolutional neural network
Data fusion
Model fusion

ABSTRACT

Purpose: Chronic obstructive pulmonary disease (COPD) is one of the most common chronic illnesses in the world. Unfortunately, COPD is often difficult to diagnose early when interventions can alter the disease course, and it is underdiagnosed or only diagnosed too late for effective treatment. Currently, spirometry is the gold standard for diagnosing COPD but it can be challenging to obtain, especially in resource-poor countries. Chest X-rays (CXRs), however, are readily available and may have the potential as a screening tool to identify patients with COPD who should undergo further testing or intervention. In this study, we used three CXR datasets alongside their respective electronic health records (EHR) to develop and externally validate our models.

Method: To leverage the performance of convolutional neural network models, we proposed two fusion schemes: (1) model-level fusion, using Bootstrap aggregating to aggregate predictions from two models, (2) data-level fusion, using CXR image data from different institutions or multi-modal data, CXR image data, and EHR data for model training. Fairness analysis was then performed to evaluate the models across different demographic groups.

Results: Our results demonstrate that DL models can detect COPD using CXRs with an area under the curve of over 0.75, which could facilitate patient screening for COPD, especially in low-resource regions where CXRs are more accessible than spirometry.

Conclusions: By using a ubiquitous test, future research could build on this work to detect COPD in patients early who would not otherwise have been diagnosed or treated, altering the course of this highly morbid disease.

1. Introduction

Chronic obstructive pulmonary disease (COPD) remains one of the leading causes of mortality and morbidity [1]. Early intervention can mitigate disease progression and decrease healthcare expenditure [2,3]. However, due to the lack of routine screening, COPD is still widely

Abbreviations: COPD, Chronic obstructive pulmonary disease; PFT, Pulmonary function tests; CXR, Chest X-ray; EHR, Electronic health records; DL, Deep learning; CNN, Convolutional neural network; MIMIC-IV, Medical Information Mart for Intensive Care; TPR, True positive rate; AUC, Area under the receiver operating characteristic curve; SD, Standard deviation; ED, Emergency department; ICU, Intensive care units; GradCAM, Gradient-weighted Class Activation Mapping.

* Corresponding author. Address: 101, Section 2, Kuang-Fu Road, Hsinchu 300044, Taiwan, ROC.
E-mail address: kuopc@cs.nthu.edu.tw (P.-C. Kuo).

https://doi.org/10.1016/j.ijmedinf.2023.105211
Received 13 June 2023; Received in revised form 23 July 2023; Accepted 1 September 2023
Available online 2 September 2023

underdiagnosed [4]. In clinical practice, pulmonary function tests (PFT) and spirometry tests are necessary diagnostic tools for COPD diagnosis [5]. Nevertheless, spirometry tests may fail to capture the presence of early COPD before it becomes detectable. Therefore, asymptomatic patients are less likely to be tested [6,7,8]. Further, the current screening tools, such as spirometry, are expensive and scarce in low-to-middle-income countries, leading to a delay in diagnosis [9,10]. Instead, the chest radiograph (CXR) is ubiquitous and cheap [11]. The ability to create an early diagnostic screening tool for COPD from CXRs offers a significant clinical opportunity in developing a tool for COPD detection and thus targeting individuals for early interventions such as medical therapy and smoking cessation programs.

Recent studies have demonstrated the ability of deep learning (DL) models, such as convolutional neural networks (CNN), to classify radiographic findings from CXRs and achieve human-like performance [12,13,14,15], while also processing large amounts of images at high speed. Previous studies have mostly focused on classifying and detecting the 14 common chest radiographic findings seen on CXRs obtained from radiology reports using publicly sourced databases [13,16,17,18]. Despite the success of DL in pulmonary disease classification using CXRs, a very limited number of studies have explored the potential of DL techniques in COPD diagnosis using CXRs only.

To the best of our knowledge, there is little attempt to develop a COPD prognostic model to identify the presence of COPD using CXR and electronic health records (EHR) data. In our study, we proposed to build deep learning techniques using model-level and data-level fusion strategies to diagnose COPD with multi-site data. Identifying COPD from CXRs and EHRs for COPD screening could prompt earlier diagnosis, prevention, and treatment.

1.1. Related work

1.1.1. COPD diagnostic models

DL strategies, such as CNN classifiers, have become a dominant approach in classifying radiographic findings from CXRs. Recent studies have explored the use of various DL techniques to predict COPD; however, these studies present several limitations. They are limited to a very small number of COPD patients in both the training and testing set, only one data type [19], and single institution training data [20,21]. Further, those studies are limited in their consideration of the mechanical ventilation status of patients and thus pose uncertainty about their utility to screen patients with COPD as outpatients.

DeGrave et al. [22] demonstrated that the artificial intelligence system would learn spurious features as shortcuts to aid their classification rather than learning pathologically relevant features. For this concern, in Schroeder et al. [20], intubation may become a shortcut for models to determine the stage of COPD because the intubation would be obvious in the CXRs and indicative of respiratory status. To our knowledge, the latent ability to recognize patients with early COPD and without mechanical ventilation solely depending on CXRs has not yet been investigated.

1.1.2. Model-level and data-level fusion

Two types of fusion strategies have been shown to be promising in prediction tasks. Model-level fusion is one approach that combines base-model predictions to form a composite prediction [23,24,25,26,27] and thus leverages the strength of each model. Bagging and stacking are two popular model-fusion techniques [25]. Bagging integrates a set of predictions from several independent base models into a single prediction using an averaging or majority voting scheme. Stacking is an approach where the meta-model learns how to combine the output of the base models best to provide an optimal prediction. Model-level fusion schemes have been implemented in different medical imaging tasks and have been shown to perform better and more stable than single-model architectures [27,28,29,30,31,32].

Data-level fusion schemes integrate multi-modal features or multi-site data into a single predictor [32,33,34]. Various studies have demonstrated consistent improvement in model performance using data-level fusion strategies [33,34,35,36]. With the proliferation of publicly available medical imaging datasets and EHR, it is important to utilize pixel-level data and clinical patient data to obtain better feature representations during training. It is also important to use data from multiple centers to account for variability in data acquisition and processing to increase the validity of the proposed fusion method [37].

1.1.3. Fairness evaluation across different subgroups

Fairness is a rising concern in DL applications in the medical field. Seyyed-Kalantari et al. [38] demonstrated that state-of-the-art DL models consistently underdiagnosed under-served patients. In Seyyed-Kalantari et al. [39]'s study, the researchers found disparities in true positive rate (TPR) for pulmonary disease diagnosis across different demographic subgroups. Because TPR requires a binary threshold which would significantly affect the calculation of TPR, we use the area under the receiver operating characteristic curve (AUC) as our fairness evaluation metric instead, which would not be affected by the threshold and can present the overall performance.

2. Material and methods

2.1. Study objective

In this project, our objective is to develop an early-screening tool for predicting COPD from non-ventilated patients. We used three large CXR datasets and one EHR dataset from CheXpert [40], Medical Information Mart for Intensive Care (MIMIC-IV) [41,42], MIMIC-CXR-JPG [43,44], and Emory-CXR to perform our experiment. CheXpert was used as a pre-train task for the model to learn the radiological features before transfer learning. MIMIC-CXR was used to fine-tune the model to classify the occurrence of COPD from CXR. The Emory-CXR was used to validate our models externally, and we also combined the Emory-CXR with MIMIC-CXR as a multi-site data level fusion. We then investigate if the following strategies can improve COPD prediction performance: (1) bagging predictions from four CNN models using CXR image data, (2) incorporating CXR image data and EHR data for modeling, and (3) using multi-site data for model training.

2.2. Data

2.2.1. Chest X-ray datasets

This study uses frontal CXRs from three datasets: CheXpert, MIMIC-CXR-JPG, and Emory-CXR, an Emory CXR dataset. CheXpert is a retrospective dataset from Stanford Hospital consisting of 224,316 CXRs of 65,240 patients [40]. MIMIC- CXR-JPG is a publicly available dataset containing 377,110 JPG images and its structured labels corresponding to 227,835 radiographic studies sourced from the Beth Israel Deaconess Medical Center between 2011 and 2016 [41,43,44]. Both CheXpert and MIMIC-CXR are labeled with the same 14 cardiopulmonary disease labels. Emory-CXR is collected from five hospitals across the Emory healthcare system between 2019 and 2020. This dataset is acquired from inpatient and outpatient hospitals and contains 226,640 CXRs from 90,483 patients.

2.2.2. EHR databases

We use the MIMIC-IV database (version: 1.0), which is a publicly available EHR database maintained by Beth Israel Deaconess Medical Center from 2008 to 2019, consisting of more than 200,000 emergency department (ED) admissions and more than 60,000 intensive care units (ICU) stays [41,42]. We also use the Emory EHR from the Emory database, collected from five hospitals across the Emory healthcare system.

2.2.3. Patient cohort and data pre-processing

We extract the patient cohort with COPD and those without COPD

from the ED and ICU admissions in the MIMIC-IV database and Emory-CXR. We create binary labels for the CXRs according to the ICD-9 and ICD-10 diagnosis codes listed in Supplementary Table A to indicate the presence or absence of COPD. We used images from patients on room air, oxygen, or high-flow nasal cannula. We excluded images from patients who were on mechanical ventilation for two reasons. First, we did not want the algorithm to recognize the mechanical ventilation device and use it as a shortcut to classify COPD. Second, our focus is to build a potential outpatient screening tool for patients with COPD, ideally for early detection, so we exclude COPD patients who are on mechanical ventilation as they are more likely to have advanced disease.

The patient's demographic obtained from MIMIC-IV and the Emory EHR data includes self-reported race-ethnicity, sex, and age as shown in Table 1. The summary of the datasets we use in our experiments is shown in Table 2. All images are processed with histogram equalization [44,45], resized to (256, 256). The image pixel values are normalized from 0 to 1. We split the imaging data into training (64%), validation (16%), and testing (20%) based on the patient's ID, and we ensure that patients in the training set are not included in the validation and testing set to avoid data leakage.

2.3. Transfer learning

In this study, we use four CNN-derived state-of-the-art models, including DenseNet121 [46], ResNet50V2 [47], MobileNetV2 [48], and Xception [49] equipped with ImageNet pre-trained weights. Since we only have a limited number of images from our cohort, the models are pre-trained on the CheXpert dataset to learn radiological features before being fine-tuned on the COPD prediction task.

In the pre-training stage, we select the top six radiographic labels, including 'Atelectasisa', 'Cardiomegaly', 'Edema', 'Lung Opacity', 'Pleural Effusion', and 'Support Devices', out of fourteen common labels based on their prevalence in the CheXpert dataset as the pre-train task to allow the models to learn pulmonary features. In the fine-tuning stage, each pre-trained model is fine-tuned on the MIMIC-CXR dataset for the COPD detection task. To accelerate the fine-tuning process, we freeze the learned weights for the first 30% of the layers and fine-tune the remaining layers. To address the imbalance in the number of cases of COPD, we use the class weights parameter to give different weightage to the COPD and non-COPD classes, which are defined as follows:

$$w_{COPD} = \frac{Total\ \#\ of\ CXRs}{2(\#\ of\ COPDCXRs)} \tag{1}$$

$$w_{Non-COPD} = \frac{Total\ \#\ of\ CXRs}{2(\#\ of\ Non-COPDCXRs)} \tag{2}$$

For the pre-training and fine-tuning stages, the learning rate is set to 0.001 and decays 5% for every epoch, and the batch size is set to 32. Adam optimizer is used, and binary cross-entropy is used as the loss function.

Table 2
Summary of the datasets used for the COPD prediction experiments.

	MIMIC	Emory	CheXpert
Number of patients (Number of images)	52,804 (194,748)	90,483 (226,640)	48,285 (140,921)
Number of COPD patients (Number of COPD images)	8,105 (50,757)	10,353 (40,793)	N/A
Number of non-COPD patients (Number of non-COPD images)	44,699 (143,991)	80,130 (185,847)	N/A

2.4. Fusion strategies

To improve the performance of the fine-tuned models, we use the fine-tuned models as the base models to implement model-level fusion and data-level fusion strategies. In terms of model-level fusion, we implement bagging methods, including unweighted and weighted averaging. For data-level fusion, we experiment with two different techniques. First, we merge the MIMIC-CXR and Emory-CXR to create a multi-site data fusion strategy, which we use as our training set to train the best-performing base model (Xception). Second, we implement a multi-modal data fusion strategy (joint fusion type II) [32], where we concatenate the feature representations of the CXRs from the penultimate Xception model layer with demographic variables from EHR. We intend to evaluate if using multiple sources of data and different modalities can enhance the performance of the base model. Our proposed approaches are shown in Fig. 1.

2.4.1. Model-level fusion

Unweighted average bagging We average the predictions of the four fine-tuned base models with equal weighting to obtain the final predictions. We later evaluate whether the unweighted average bagging strategy outperforms the single-base models.

Weighted average bagging We incorporated the concept of hierarchical clustering to implement the weighted average bagging technique in our study. That is, we average the test predictions of the base models based on the distance of the validation set predictions dervied from the four base models to obtain the final prediction. The implementation details regarding the calculation of the distance and the averaging approach are shown in Appendix C. We calculate the cosine distance of the validation set predictions obtained from the four base models and construct a dendrogram using the Unweighted Pair Group Method with an Arithmetic mean (UPGMA) algorithm, a hierarchical clustering approach [50] defined as follows:

Table 1
Summary of the demographic information of patients in MIMIC and Emory datasets.

	MIMIC			Emory		
	COPD	Non-COPD	P-Value	COPD	Non-COPD	P-Value
# of patients	8,105	44,699		10,353	10,353	
Asian	175 (2.2%)	1,755 (3.9%)	<0.001	239 (2.3%)	3,054 (3.8%)	<0.001
Black	1,162 (14.3%)	7,761 (17.4%)		4,313 (41.7%)	38,047 (47.5%)	
Latino	321 (4.0%)	3,019 (6.8%)		125 (1.2%)	1,447 (1.8%)	
Others	584 (7.2%)	4,136 (9.3%)		341 (3.3%)	4,863 (6.1%)	
White	5,863 (72.3%)	28,028 (62.7%)		5,335 (51.5%)	32,719 (40.8%)	
Female	4,146 (51.2%)	23,327 (52.2%)	0.089	4,918 (47.5%)	43,539 (54.3%)	<0.001
Male	3,959 (48.8%)	21,372 (47.8%)		5,435 (52.5%)	36,591 (45.7%)	
0–40	208 (2.6%)	8,008 (17.9%)	<0.001	310 (3.0%)	19,234 (24.0%)	<0.001
40–60	1,975 (24.4%)	14,458 (32.3%)		1,827 (18.1%)	25,037 (31.2%)	
60–80	4,093 (50.5%)	15,400 (34.5%)		5,957 (57.5%)	27,123 (33.8%)	
80+	1829 (22.6%)	6,833 (15.3%)		2,214 (21.4%)	8,736 (10.9%)	

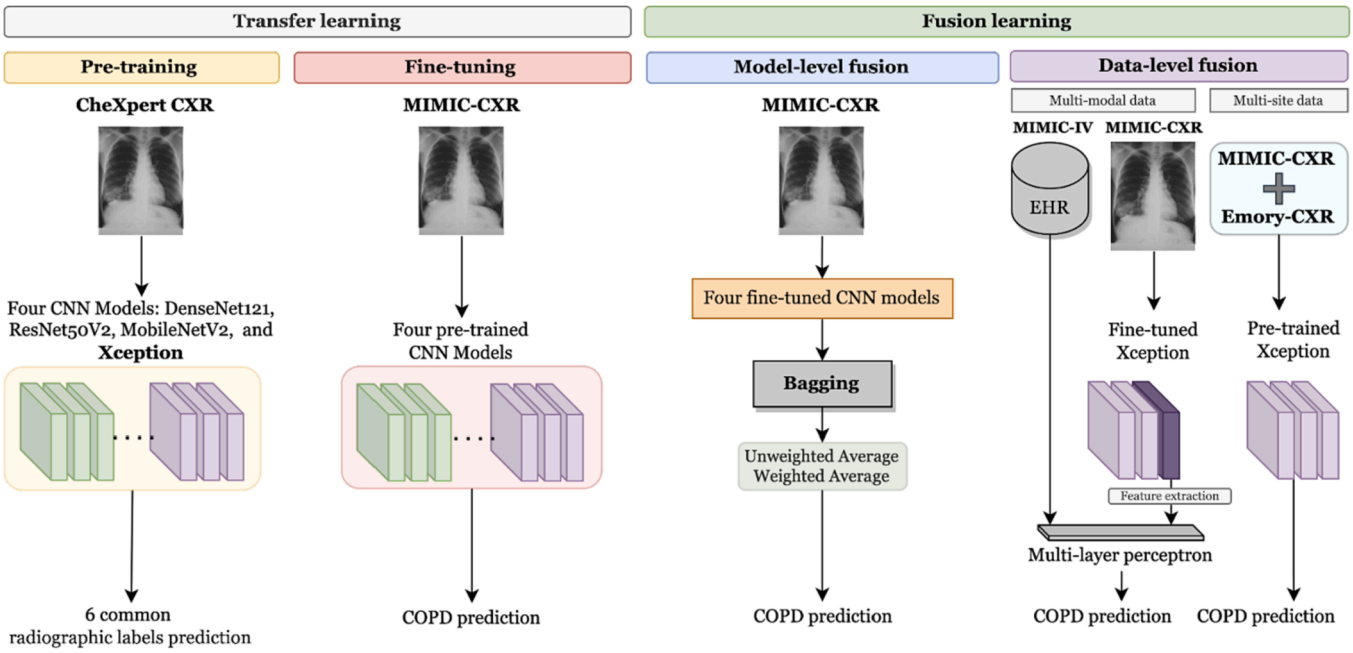


Fig. 1. Overview of our proposed framework.

$$d(U,V) = \sum_{i}^{i\in U}\sum_{j}^{j\in V}\frac{d(U_i,V_j)}{|U||V|} \tag{3}$$

where |U| and |V| are the cardinalities of clusters U and V, respectively for all points i and j. We average the test predictions of the base models based on the dendrogram as shown in Fig. 2 to obtain the final prediction. The dendrogram provides an intuitive and effective way to visualize the intricate relationships between models' predictions. For example, in Fig. 2, the predictions of the DenseNet121 and the ResNet50V2 are averaged first, followed by the predictions of the MobileNetV2, and finally, the predictions of the Xception. This is equivalent to assigning weights of the predictions of the DenseNet121, ResNet50V2, MobileNetV2, and Xception as 1/8, 1/8, 1/4, and 1/2, respectively. We give the prediction of the base model a larger weight since the length of the dendrogram leg for the base model is longer.

2.4.2. *Data-level fusion*

Multi-site data fusion We merge two datasets, MIMIC-CXR and Emory-CXR, and use the combined dataset to fine-tune and evaluate the best-performing base model (Xception). We examine if data from different sources enhances the performance and generalizability of the multi-site data fusion strategy.

Multi-modal data fusion We construct a data-level fusion strategy with the CXRs and their corresponding EHRs as input. We use an

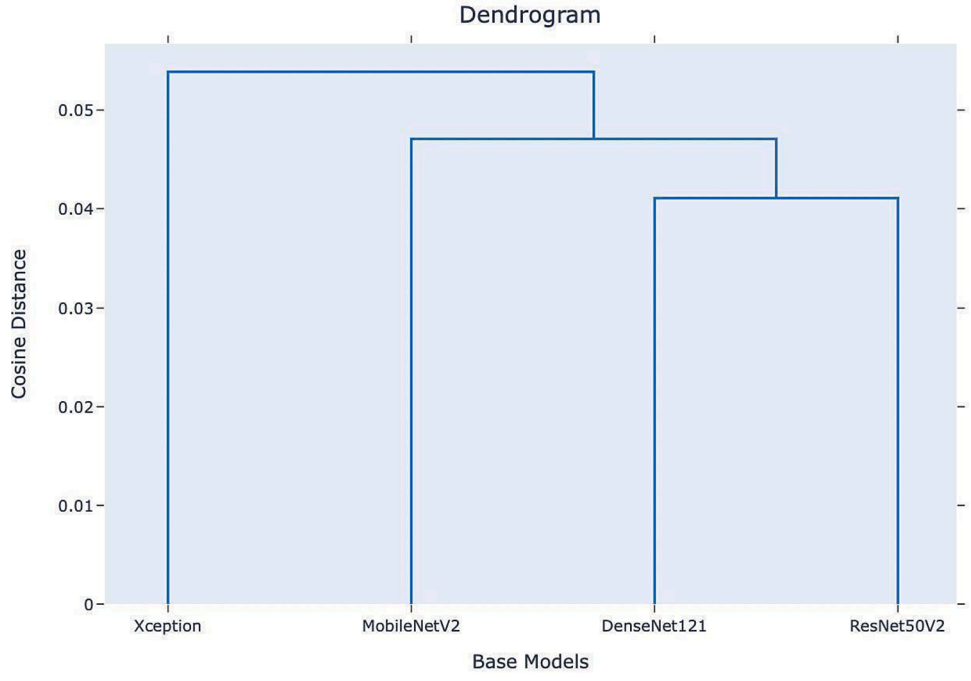


Fig. 2. Dendrogram showing pairwise cosine distance for predictions across the four base models.

Xception base model and a multi-layer perceptron model to create a multi-modal strategy (joint fusion—type II) [32]. The learned features representations of the CXRs from the penultimate fine-tuned Xception layer are encoded into 64 hidden nodes concatenated with ten input nodes representing the demographic variables —five for race-ethnicity (Asian, Black, Hispanic, Others, and White), four for age (0–40, 40–60, 60–80, and 80+), and one for sex (Female or Male).

2.5. Evaluation metrics

We report the point estimates and the 95% confidence interval (CI) of the evaluation metrics, which is calculated by bootstrapping the metric scores over 1000 runs: AUC, recall, and F1-score for all the experimented and proposed methodologies. AUC is the most used evaluation metric when it comes to evaluating the overall performance of the model. In the early-screening situation, we need to focus on recall. The binary threshold for recall and F1-score is calculated to maximize the F1-score for the validation set. We evaluate our model using internal test data (MIMIC-CXR), and external test data (Emory-CXR).

2.6. Model interpretation

To understand which CXR regions the base models focus on, we visualize the hot spots for the true positive by using the Gradient-weighted Class Activation Mapping (Grad-CAM) algorithm [51].

2.7. Fairness analysis

To examine whether there is a discrepancy in the models' performance across demographic subgroups, we evaluate the average AUC and the standard deviation (SD) of the AUCs among different race-ethnicity, age groups, and sex separately on the MIMIC-CXR and Emory-CXR. That is, the model would have separate results for each demographic group. The average AUC shows how the models perform in the demographic group, and the SD shows the deviation from the average AUC across demographic subgroups.

3. Results

3.1. Transfer learning

We present the results of four CNN models tested on Emory-CXR in Table 3. The Xception base model yields the best AUC score (0.74), F1-score (0.43), and Recall (0.72) during the external testing on Emory-CXR. We conducted DeLong [52] tests to examine whether there were significant differences in the AUCs among the four base models. The results indicated that the AUC values for Xception were significantly higher than those for ResNet50V2 ($p < 0.01$), MobileNetV2 ($p < 0.01$), and DenseNet121 ($p < 0.01$). The internal testing result on MIMIC-CXR data and the detailed results are listed in Supplementary Table B1 and B2.

3.2. Fusion strategies

We present the external testing results of our fusion strategies in Table 4 for Emory-CXR. The Multi-site data-level fusion has the best AUC score (0.76) and F1-score (0.45). The weighted bagging has the best Recall (0.73). Since the Xception model performs the best among the base models, we use the Xception model in the fusion strategies. We also implement four machine learning models as the meta-models, including logistic regression, k-nearest neighbors, XGBoost, and random forest. The stacking model using the random forest as a meta-model performs the best among the other stacking models; however, it does not outperform the Xception base model. The results of the stacking models are listed in Supplementary Table B3-B4 and B11-B12. The internal testing result on MIMIC-CXR data and detailed results are listed in Supplementary Table B3 and B4.

Table 3
Base models tested on the Emory CXR image data.

Model	AUC	F1-score	Recall
DenseNet121	0.73 [0.72–0.73]	0.43 [0.42–0.44]	0.66 [0.65–0.67]
MobileNetV2	0.72 [0.72–0.73]	0.42 [0.41–0.43]	0.68 [0.67–0.69]
ResNet50V2	0.72 [0.71–0.73]	0.42 [0.41–0.43]	0.67 [0.66–0.68]
Xception	0.74 [0.73–0.75]	0.43 [0.43–0.44]	0.72 [0.72–0.73]

Table 4
Evaluation metrics averaged over 1000 epochs ± 95% CI for the various fusion strategies on the Emory test data compared to the Xception base model.

	AUC	F1-score	Recall
Fusion Strategies			
Base Xception Model	0.74 [0.73–0.75]	0.43 [0.43–0.44]	0.72 [0.72–0.73]
Base Models Bagging			
Unweighted Average	0.75 [0.74–0.75]	0.44 [0.43–0.45]	0.71 [0.70–0.72]
Weighted Average	0.75 [0.74–0.75]	0.44 [0.43–0.45]	0.73 [0.72–0.74]
Data-level Fusion			
Multi-site	0.76 [0.75–0.76]	0.45 [0.44–0.45]	0.72 [0.71–0.73]
Multi-modal	0.74 [0.73–0.74]	0.44 [0.43–0.45]	0.67 [0.66–0.69]

The multi-site data-level fusion strategy performs better when testing on Emory-CXR than the Xception base model, which shows that combining multi-site data can improve performance. The multi-modal data-level fusion strategy results show that incorporating demographic features does not improve performance. We use the chi-square test to evaluate the independence of the race-ethnicity, age, sex variables, and the 14 radiographic labels between the COPD and non-COPD cohorts. A lower p-value ($p < 0.01$) demonstrates that the distributions of these variables are statistically significant between the COPD and non-COPD patient cohorts.

In addition, we employed a logistic regression model to assess the significance of various features, including image, race, age, and gender. By evaluating the coefficients of the logistic regression model, we obtained the feature importance. Fig. 3 illustrates that image features are the most influential, followed by age, race, and gender, in descending order of importance. This outcome indicates that even with the incorporation of EHR data, image features remain the primary driver of predictive power. Consequently, the combination of EHR data did not lead to a notable improvement in the overall performance of the model. This finding underscores the continued dominance of image features in accurately predicting COPD cases.

3.3. Explainability with Grad-CAM

We visualized the Grad-CAM heatmaps of the true positive cases on the Xception base model for both MIMIC-CXR and the Emory test datasets (Emory-CXR). Fig. 4 shows the heatmaps of the true positive cases of the Xception model. We further average the heatmaps with predicted probabilities higher than 0.9 to obtain an averaged heatmap (Mean). We can see that the Xception base model mainly focuses on the apical lung fields, which can be consistent with centrilobular emphysema, a common form of COPD affecting the upper lobe lung fields [53,54]. This is likely why the models are identifying these regions to make predictions.

3.4. Fairness analysis

We evaluate the performance of the fusion strategies on each demographic group separately in both datasets. Table 5 shows the results of the fairness analysis of each model on Emory-CXR. The average AUC

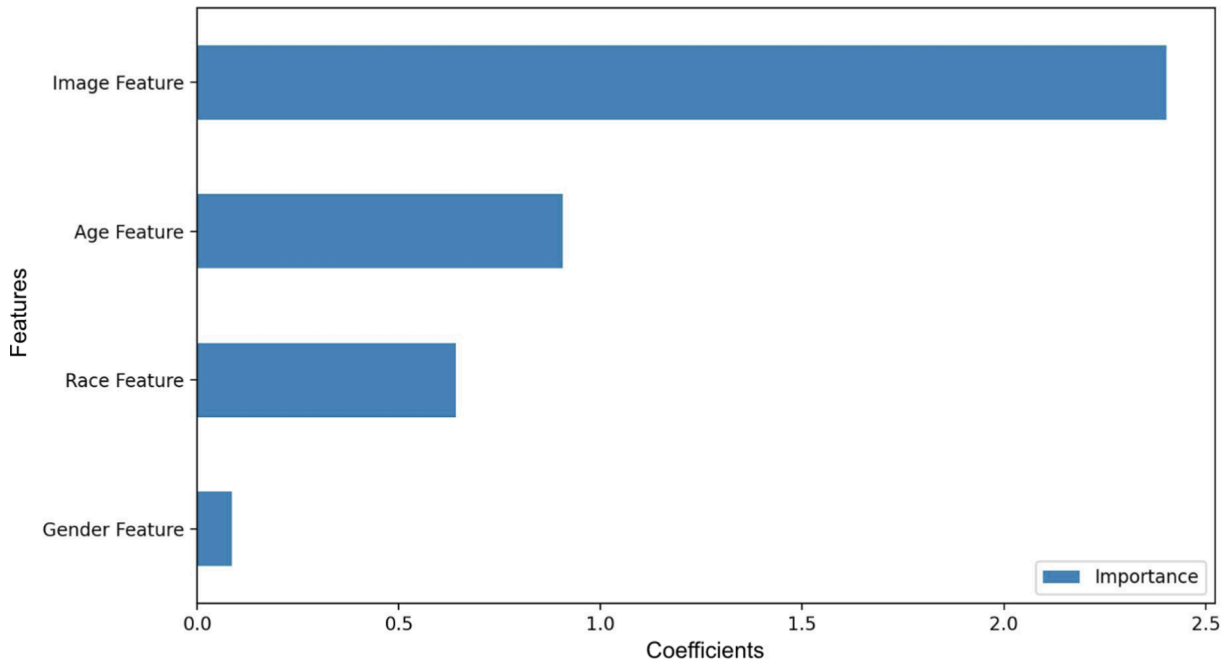


Fig. 3. The feature importance of the multi-modal data-level fusion obtained from coefficients of the logistic regression model.

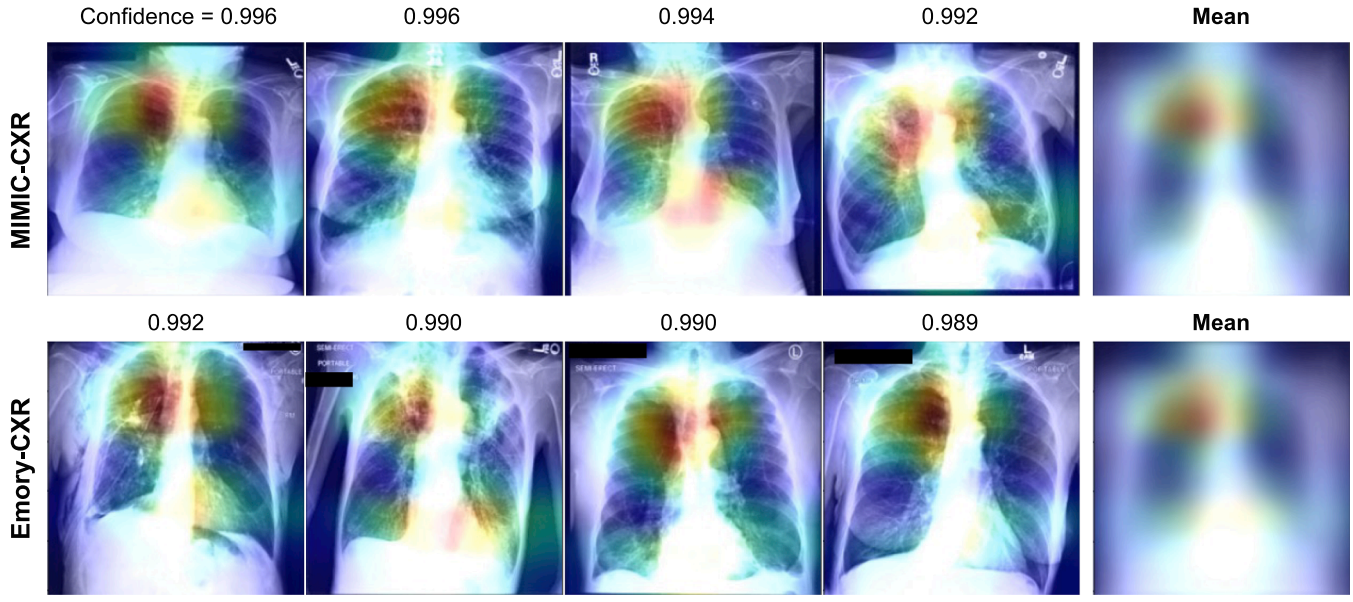


Fig. 4. The first and second rows display the Grad-CAM heatmaps of the true positive cases on the Xception base model using the MIMIC-CXR test data and the Emory CXR image test data, respectively.

scores show the average performance of the models among each demographic group, and the standard deviations show the performance gap between each demographic group. The star signs in the table represent the highest standard deviation, indicating the model's unfairness. The baseline and weighted average bagging models are the only two that do not have a star sign. That is, both models are reasonably fair across demographic groups. The detailed results with precision and recall are listed in Supplementary Table B5 to B16.

4. Discussion

Our findings support our hypothesis that DL models trained using CXRs can identify patients with COPD. CXRs are a more accessible and affordable modality. Therefore, our approach could provide a low-cost and widely available screening opportunity for COPD. We demonstrate that the Xception base model detects COPD with reasonable performance compared to the other base models, as shown in Table 3. The results are consistent for both MIMIC-CXR and the Emory-CXR test datasets across the fusion strategies. The multi-site data-level fusion scheme performs better than the other schemes on the Emory test data indicating that our proposed strategy is generalizable. We use Grad-CAM to evaluate the Xception model's explainability. Looking at the true positive cases for both MIMIC-CXR and Emory-CXR, we can see that the Grad-CAM heatmap of the Xception base model focuses on the upper lung fields.

The limitations of this study include reliance on ICD9/10 coding as

Table 5
Average AUC ± SD for model-level and data-level fusion strategies on Emory-CXR across various demographic subgroups.

Fusion Strategies	Race-ethnicity	Sex	Age
Base Xception Model	0.72 [0.024]	0.74 [0]	0.70 [0.030]
Unweighted Average Bagging	0.72 [0.038]	0.75 [0.005] *	0.71 [0.025]
Weighted Average Bagging	0.73 [0.030]	0.75 [0]	0.71 [0.027]
Base Models Stacking (Random Forest Classifier)	0.71 [0.038]	0.74 [0.005] *	0.70 [0.029]
Multi-site Data Fusion	0.74 [0.050] *	0.76 [0]	0.72 [0.031] *
Multi-modal Data Fusion	0.72 [0.038]	0.74 [0]	0.69 [0.026]

labels and the lack of Pulmonary Function Testing (PFT) labels. PFT is an effective tool for the diagnosis of COPD. However, our study uses ICD codes to identify COPD diagnosis because concurrent PFT data were unavailable for many of the CXRs in our datasets. Our study proves that deep learning can be applied to patient data even if gold-standard PFT findings are unavailable. This has widespread implications because the current diagnostic criteria for COPD lead to a large population of underdiagnosed patients [55]. Additionally, clinicians often only order PFTs when COPD progresses and becomes symptomatic, indicating that many early-stage diagnoses are often missed [56]. Another limitation of this study is the unavailability of information regarding the stage of COPD for the patients in the current datasets. A prospective study is essential to assess the algorithm's performance across different stages of COPD. However, it is important to note that the intended use of this algorithm does not involve screening asymptomatic patients. Instead, the objective is to leverage CXRs, a widely accessible test obtained for various reasons in health centers, clinics, and hospitals, to detect COPD, whether in its early or advanced stages. This approach is particularly valuable for health systems with limited resources, as it enables the identification of the disease even before it reaches an irreversible and advanced state.

The limitation of model explainability should be noted in this study. While Grad-CAM is commonly used for visualization, it comes with certain drawbacks, such as uncertainty, lack of robustness, and limited usefulness in medical imaging analysis. Studies like Chattopadhyay et al. [57] have demonstrated that the Grad-CAM method can fail to precisely localize the object of interest and might struggle with locating multi-occurrence objects effectively. Moreover, research by Demir et al. [58] has highlighted the lack of robustness in Grad-CAM, as its gradient-based approach makes it sensitive to various parameters within the network. To address these concerns, we chose to present the average heatmap across images instead of relying on a single case for visualization. This approach was based on feedback from physicians, who found the average heatmap to be more meaningful and informative.

Our work is the first step to depicting the potential of deep learning in detecting COPD patients. Given COPD is one of the largest killers worldwide, especially in resource-poor countries, the ultimate goal would be to have a model that can screen patients for COPD who are undergoing an easily accessible CXR for any reason, and identify those who are high-risk and should seek either further testing with spirometry or empiric intervention such as smoking cessation or medical therapy. This has the potential to alter the disease course, saving patients and hospital systems from the advanced stages of this underdiagnosed, burdensome disease. We encourage other hospitals to similarly build their own early COPD screening model using their databases based on our open-source code.

5. Conclusion

Our proposed fusion schemes, along with the fairness analysis, provide a framework for future research to build upon. By using a widely available test like CXRs, we may be able to diagnose COPD earlier and improve outcomes for patients with this highly morbid disease.

Summary table

Background

- COPD is one of the most common chronic illnesses in the world and is often underdiagnosed or only diagnosed too late for effective treatment.
- Current COPD diagnoses rely on spirometry which can be challenging to obtain.
- CXRs are readily available and may have the potential as a screening tool to identify patients with COPD who should undergo further testing or intervention.
- There is little attempt to develop a COPD prognostic model to identify the presence of COPD using CXR and EHR data.

Our contribution

- We implement data- and model-level fusion strategies to build a COPD screening model.
- We combine clinical data and CXRs from multiple sites to enhance the model's generalizability.
- We assess model bias across different demographic subgroups, including race-ethnicity, sex, and age.

CRediT authorship contribution statement

Ryan Wang: Conceptualization, Data curation, Methodology, Software, Validation, Writing – original draft, review & editing. Li-Ching Chen: Conceptualization, Writing – original draft. Lama Moukheiber: Conceptualization, Writing – original draft. Kenneth P. Seastedt: Conceptualization, Writing – review & editing. Mira Moukheiber: Writing – review & editing. Dana Moukheiber: Writing – review & editing. Zachary Zaiman: Data curation, Software. Sulaiman Moukheiber: Writing – review & editing. Tess Litchman: Conceptualization. Hari Trivedi: Resources. Rebecca Steinberg: Resources. Judy W. Gichoya: Conceptualization, Resources, Writing – review & editing. Po-Chih Kuo: Conceptualization, Methodology, Validation, Project administration, Resources, Supervision, Writing – original draft, review & editing. Leo A. Celi: Conceptualization, Supervision, Writing – review & editing.

CRediT authorship contribution statement

Ryan Wang: Conceptualization, Data curation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Li-Ching Chen:** Conceptualization, Writing – original draft. **Lama Moukheiber:** Conceptualization, Writing – original draft. **Kenneth P. Seastedt:** Conceptualization, Writing – review & editing. **Mira Moukheiber:** Writing – review & editing. **Dana Moukheiber:** Writing – review & editing. **Zachary Zaiman:** Data curation, Software. **Sulaiman Moukheiber:** Writing – review & editing. **Tess Litchman:** Conceptualization. **Hari Trivedi:** Resources. **Rebecca Steinberg:** Resources. **Judy W. Gichoya:** Conceptualization, Resources, Writing – review & editing. **Po-Chih Kuo:** Conceptualization, Methodology, Validation, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing. **Leo A. Celi:** Conceptualization, Supervision, Writing – review & editing.

Funding

This study is funded by the National Science and Technology Council, Taiwan (MOST109-2222-E-007-004-MY3). L.A.C and D.M. are

funded by the National Institute of Health through the NIBIB R01 EB017205. D.M. and L.M. are supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under NIH grant number R01EB030362. D.M. is supported by NIH National Library of Medicine under contract number 75N97020C00013, and Massachusetts Life Sciences Center, Jul. 1st, 2020. J.W.G. declares support from US National Science Foundation (grant number 1928481) from the Division of Electrical, Communication & Cyber Systems, RSNA Health Disparities grant (#EIHD2204), NIH (NIBIB) MIDRC grant under contracts 75N92020C00008 and 75N92020C00021.

Availability of data and materials

The datasets supporting the conclusions of this article are available at: https://physionet.org/. All codes in Python for the study are available at: https://github.com/Ryan-RE-Wang/COPD-project/.

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2023.105211.

References

[1] D.M. Mannino, A.S. Buist, Global burden of COPD: Risk factors, prevalence, and future trends, Lancet 370 (9589) (2007) 765–773.
[2] C. Moretz, Y. Zhou, A.D. Dhamane, K. Burslem, K. Saverno, G. Jain, G. Devercelli, S. Kaila, J.J. Ellis, G. Hernandez, A. Renda, Development and validation of a predictive model to identify individuals likely to have undiagnosed chronic obstructive pulmonary disease using an administrative claims database, J. Manag. Care Spec. Pharm. 21 (12) (2015) 1149–1159.
[3] M. Decramer, C.B. Cooper, Treatment of COPD: the sooner the better? Thorax 65 (9) (2010) 837–841.
[4] K. Hill, R.S. Goldstein, G.H. Guyatt, M. Blouin, W.C. Tan, L.L. Davis, D.M. Heels-Ansdell, M. Erak, P.J. Bragaglia, I.E. Tamari, R. Hodder, M.B. Stanbrook, Prevalence and underdiagnosis of chronic obstructive pulmonary disease among patients at risk in primary care, Can. Med. Assoc. J. 182 (7) (2010) 673–678.
[5] C.F. Vogelmeier, G.J. Criner, F.J. Martinez, A. Anzueto, P.J. Barnes, J. Bourbeau, B. R. Celli, R. Chen, M. Decramer, L.M. Fabbri, P. Frith, D.M.G. Halpin, M.V. López Varela, M. Nishimura, N. Roche, R. Rodriguez-Roisin, D.D. Sin, D. Singh, R. Stockley, J. Vestbo, J.A. Wedzicha, A. Agusti, Global strategy for the diagnosis, management and Prevention of Chronic Obstructive Lung Disease 2017 report, Respirology 22 (3) (2017) 575–601.
[6] L.E. Labonté, W.C. Tan, P.Z. Li, P. Mancino, S.D. Aaron, A. Benedetti, K. R. Chapman, R. Cowie, J.M. FitzGerald, P. Hernandez, F. Maltais, D.D. Marciniuk, D. O'Donnell, D. Sin, J. Bourbeau, Undiagnosed chronic obstructive pulmonary disease contributes to the burden of health care use. Data from the Cancold Study, Am. J. Respir. Crit. Care Med. 194 (3) (2016) 285–298.
[7] A.L. Siu, K. Bibbins-Domingo, D.C. Grossman, K.W. Davidson, J.W. Epling, F. A. García, M. Gillman, A.R. Kemper, A.H. Krist, A.E. Kurth, C.S. Landefeld, C. M. Mangione, D.M. Harper, W.R. Phillips, M.G. Phipps, M.P. Pignone, Screening for chronic obstructive pulmonary disease, J. Am. Med. Assoc. 315 (13) (2016) 1372.
[8] E. Andreeva, M. Pokhaznikova, A. Lebedev, I. Moiseeva, O. Kuznetsova, J.-M. Degryse, Spirometry is not enough to diagnose COPD in epidemiological studies: a follow-up study, npj Primary Care Respir. Med. 27 (1) (2017) pp.
[9] D. Beran, H.J. Zar, C. Perrin, A.M. Menezes, P. Burney, Burden of asthma and chronic obstructive pulmonary disease and access to essential medicines in low-income and middle-income countries, Lancet Respir. Med. 3 (2) (2015) 159–170.
[10] J. Meghji, K. Mortimer, A. Agusti, B.W. Allwood, I. Asher, E.D. Bateman, K. Bissell, C.E. Bolton, A. Bush, B. Celli, C.-Y. Chiang, A.A. Cruz, A.-T. Dinh-Xuan, A. El Sony, K.M. Fong, P.I. Fujiwara, M. Gaga, L. Garcia-Marcos, D.M. Halpin, J.R. Hurst, S. Jayasooriya, A. Kumar, M.V. Lopez-Varela, R. Masekela, B.H. Mbatchou Ngahane, M. Montes de Oca, N. Pearce, H.K. Reddel, S. Salvi, S.J. Singh, C. Varghese, C.F. Vogelmeier, P. Walker, H.J. Zar, G.B. Marks, Improving lung health in low-income and middle-income countries: from challenges to solutions, Lancet 397 (10277) (2021) 928–940.
[11] F.A. Mettler, M. Mahesh, M. Bhargavan-Chatfield, C.E. Chambers, J.G. Elee, D. P. Frush, D.L. Miller, H.D. Royal, M.T. Milano, D.C. Spelic, A.J. Ansari, W.E. Bolch, G.M. Guebert, R.H. Sherrier, J.M. Smith, R.J. Vetter, Patient exposure from radiologic and nuclear medicine procedures in the United States: procedure volume and effective dose for the period 2006–2016, Radiology 295 (2) (2020) 418–427.
[12] P. Lakhani, B. Sundaram, Deep learning at chest radiography: Automated Classification of pulmonary tuberculosis by using convolutional neural networks, Radiology 284 (2) (2017) 574–582.
[13] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," arXiv.org, 25-Dec-2017. [Online]. http://arxiv.org/abs/1711.05225 (accessed: 22-Mar-2023).
[14] C. Wang, A. Elazab, J. Wu, Q. Hu, Lung nodule classification using deep feature fusion in chest radiography, Comput. Med. Imaging Graph. 57 (2017) 10–18.
[15] Y.-X. Tang, Y.-B. Tang, Y. Peng, K. Yan, M. Bagheri, B.A. Redd, C.J. Brandon, Z. Lu, M. Han, J. Xiao, R.M. Summers, Automated abnormality classification of chest radiographs using deep convolutional neural networks, npj Digital Med. 3 (1) (2020).
[16] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, and K. Lyman, "Learning to diagnose from scratch by exploiting dependencies among labels," arXiv.org, 01-Feb-2018. [Online] https://doi.org/10.48550/arXiv.1710.10501 (Accessed: 22-Mar-2023).
[17] 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: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[18] J. Rubin, D. Sanghavi, C. Zhao, K. Lee, A. Qadir, M. Xu-Wilson, Large scale automated reading of frontal and lateral chest x-rays using dual convolutional neural networks, arXiv.org, 24-Apr-2018. [Online] https://doi.org/10.48550/arXiv.1804.07839 (accessed: 22-Mar-2023).
[19] A. Pyrros, J. Rodriguez Fernandez, S.M. Borstelmann, A. Flanders, D. Wenzke, E. Hart, J.M. Horowitz, P. Nikolaidis, M. Willis, A. Chen, P. Cole, N. Siddiqui, M. Muzaffar, N. Muzaffar, J. McVean, M. Menchaca, A.K. Katsaggelos, S. Koyejo, W. Galanter, Validation of a deep learning, value-based care model to predict mortality and comorbidities from chest radiographs in covid-19, PLOS DigitalHealth 1 (8) (2022) pp.
[20] J. D. Schroeder, R. Bigolin Lanfredi, T. Li, J. Chan, C. Vachet, R. Paine, V. Srikumar, T. Tasdizen, Prediction of obstructive lung disease from chest radiographs via deep learning trained on pulmonary function data, Int. J. Chronic Obstruct. Pulm. Dis. 15 (2021) 3455–3466.
[21] J.G. Nam, H.-R. Kang, S.M. Lee, H. Kim, C. Rhee, J.M. Goo, Y.-M. Oh, C.-H. Lee, C. M. Park, Deep learning prediction of survival in patients with chronic obstructive pulmonary disease using chest radiographs, Radiology 305 (1) (2022) 199–208.
[22] A.J. DeGrave, J.D. Janizek, S.-I. Lee, AI for radiographic COVID-19 detection selects shortcuts over signal, Nat. Mach. Intell. 3 (7) (2021) 610–619.
[23] Y. Cao, T.A. Geddes, J.Y. Yang, P. Yang, Ensemble deep learning in bioinformatics, Nat. Mach. Intell. 2 (9) (2020) 500–508.
[24] M.A. Ganaie, M. Hu, A.K. Malik, M. Tanveer, P.N. Suganthan, Ensemble deep learning: a review, Eng. Appl. Artif. Intel. 115 (2022), 105151.
[25] M. Ragab, K. Eljaaly, N.A. Alhakamy, H.A. Alhadrami, A.A. Bahaddad, S.M. Abo-Dahab, E.M. Khalil, Deep ensemble model for covid-19 diagnosis and classification using chest CT images, Biology 11 (1) (2021) 43.
[26] 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) 1–17.
[27] H. Jia, Y. Xia, Y. Song, W. Cai, M. Fulham, D.D. Feng, Atlas registration and ensemble deep convolutional neural network-based prostate segmentation using magnetic resonance imaging, Neurocomputing 275 (2018) 1358–1369.
[28] T.B. Chandra, K. Verma, B.K. Singh, D. Jain, S.S. Netam, Coronavirus disease (COVID-19) detection in chest X-ray images using majority voting based classifier ensemble, Expert Syst. Appl. 165 (2021), 113909.
[29] A. Barragán-Montero, U. Javaid, G. Valdés, D. Nguyen, P. Desbordes, B. Macq, S. Willems, L. Vandewinckele, M. Holmström, F. Löfman, S. Michiels, K. Souris, E. Sterpin, J.A. Lee, Artificial Intelligence and machine learning for medical imaging: a technology review, Phys. Med. 83 (2021) 242–256.
[30] E.A. Regan, D.A. Lynch, D. Curran-Everett, J.L. Curtis, J.H. Austin, P.A. Grenier, H.-U. Kauczor, W.C. Bailey, D.L. DeMeo, R.H. Casaburi, P. Friedman, E.J. Van Beek, J.E. Hokanson, R.P. Bowler, T.H. Beaty, G.R. Washko, M.L.K. Han, V. Kim, S. S. Kim, K. Yagihashi, L. Washington, C.E. McEvoy, C. Tanner, D.M. Mannino, B. J. Make, E.K. Silverman, J.D. Crapo, Clinical and radiologic disease in smokers with normal spirometry, JAMA Intern. Med. 175 (9) (2015) 1539.
[31] I. Sirazitdinov, M. Kholiavchenko, T. Mustafaev, Y. Yixuan, R. Kuleev, B. Ibragimov, Deep neural network ensemble for pneumonia localization from a large-scale chest X-ray database, Comput. Electr. Eng. 78 (2019) 388–399.
[32] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, M.P. Lungren, Fusion of medical imaging and Electronic Health Records using deep learning: a systematic review and Implementation Guidelines, npj Dig. Med. 3(1) (2020).
[33] K.-L. Du, M.N. Swamy, Combining multiple learners: data fusion and ensemble learning, Neural Netw. Stat. Learn. (2019) 737–767.
[34] S.-C. Huang, A. Pareek, R. Zamanian, I. Banerjee, M.P. Lungren, Multimodal fusion with deep neural networks for leveraging CT imaging and Electronic Health Record: a case-study in pulmonary embolism detection, Sci. Rep. 10(1) (2020).
[35] A. Tariq, L.A. Celi, J.M. Newsome, S. Purkayastha, N.K. Bhatia, H. Trivedi, J.W. Gichoya, I. Banerjee, Patient-specific COVID-19 resource utilization prediction using fusion AI model, npj Dig. Med. 4(1) (2021).
[36] R. Wang, P. Chaudhari, and C. Davatzikos, "Bias in machine learning models can be significantly mitigated by careful training: evidence from neuroimaging studies, Proc. Natl. Acad. Sci. 120(6) (2023).
[37] L. Flynn, The benefits and challenges of multisite studies: lessons learned, AACN Adv. Crit. Care 20 (4) (2009) 388–391.

[38] L. Seyyed-Kalantari, H. Zhang, M.B. McDermott, I.Y. Chen, M. Ghassemi, Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations, Nat. Med. 27 (12) (2021) 2176–2182.
[39] L. Seyyed-Kalantari, G. Liu, M. McDermott, I.Y. Chen, M. Ghassemi, Chexclusion: fairness gaps in deep chest X-ray classifiers, Biocomputing (2021, 2020.).
[40] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D.A. Mong, S.S. Halabi, J. K. Sandberg, R. Jones, D.B. Larson, C.P. Langlotz, B.N. Patel, M.P. Lungren, A. Y. Ng, Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison, Proc. AAAI Conf. Artif. Intell. 33 (01) (2019) 590–597.
[41] A.L. Goldberger, L.A. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, H.E. Stanley, Physiobank, PhysioToolkit, and PhysioNet, Circulation 101(23) (2000).
[42] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L.A. Celi, R. Mark, "Mimic-IV," MIMIC-IV v1.0, 16-Mar-2021. [Online]. https://physionet.org/content/mimiciv/1.0/ (accessed: 22-Mar-2023).
[43] A. E. Johnson, T.J. Pollard, S.J. Berkowitz, N.R. Greenbaum, M.P. Lungren, C.-ying Deng, R.G. Mark, S. Horng, "Mimic-CXR, a de-identified publicly available database of chest radiographs with free-text reports, Scientific Data 6(1) (2019).
[44] A.E.W. Johnson, T.J. Pollard, N.R. Greenbaum, M.P. Lungren, C.-ying Deng, Y. Peng, Z. Lu, R.G. Mark, S.J. Berkowitz, S. Horng, "Mimic-CXR-JPG, a large publicly available database of labeled chest radiographs," arXiv.org, 14-Nov-2019. [Online] https://arxiv.org/abs/1901.07042 (Accessed: 22-Mar-2023).
[45] G. Bradski, The OpenCV Library, Dr Dobb's J. Softw. Tools (2000).
[46] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely Connected Convolutional Networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261–2269, https://doi.org/10.1109/CVPR.2017.243.
[47] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.
[48] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: inverted residuals and linear bottlenecks, Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (Jun 2018) 4510–4520.
[49] F. Chollet, Xception: deep learning with depthwise separable convolutions, in: In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.
[50] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. van der Walt, M. Brett, J. Wilson, K.J. Millman, N. Mayorov, A.R. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E.W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E.A. Quintero, C.R. Harris, A.M. Archibald, A.H. Ribeiro, F. Pedregosa, P. van Mulbregt, A. Vijaykumar, A.P. Bardelli, A. Rothberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C.N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D.A. Nicholson, D.R. Hagen, D. V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G.A. Price, G.-L. Ingold, G.E. Allen, G.R. Lee, H. Audren, I. Probst, J. P. Dietrich, J. Silterra, J.T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J. L. Schönberger, J.V. de Miranda Cardoso, J. Reimer, J. Harrington, J.L. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N.J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P.A. Brodtkorb, P. Lee, R.T. McGibbon, R. Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T.J. Pingel, T. P. Robitaille, T. Spura, T.R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y.O. Halchenko, Y. Vázquez-Baeza, SciPy 1.0: Fundamental algorithms for scientific computing in python, Nat. Methods 17 (3) (2020) 261–272.
[51] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization. 2017 IEEE International Conference on Computer Vision (ICCV), 2017.
[52] E.R. DeLong, et al., Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics 44 (3) (1988) 837–845.
[53] S.F. Nemec, A.A. Bankier, R.L. Eisenberg, Lower lobe—predominant diseases of the lung, Am. J. Roentgenol. 200 (4) (2013) 712–728.
[54] J.R. Hurst, Upper Airway. 3: Sinonasal involvement in chronic obstructive pulmonary disease, Thorax 65 (1) (2009) 85–90.
[55] P.G. Woodruff, R.G. Barr, E. Bleecker, S.A. Christenson, D. Couper, J.L. Curtis, N. A. Gouskova, N.N. Hansel, E.A. Hoffman, R.E. Kanner, E. Kleerup, S.C. Lazarus, F. J. Martinez, R. Paine, S. Rennard, D.P. Tashkin, M.L.K. Han, Clinical significance of symptoms in smokers with preserved pulmonary function, N. Engl. J. Med. 374 (19) (2016) 1811–1821.
[56] A. Sood, H. Petersen, C. Qualls, P.M. Meek, R. Vazquez-Guillamet, B.R. Celli, Y. Tesfaigzi, Spirometric variability in smokers: transitions in COPD diagnosis in a five-year longitudinal study, Respir. Res. 17(1) (2016).
[57] A. Chattopadhay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 2018, pp. 839-847, 10.1109/WACV.2018.00097.
[58] Ugur Demir et al. Information Bottleneck Attribution for Visual Explanations of Diagnosis and Prognosis, in: Machine learning in medical imaging. MLMI (Workshop) vol. 12966 (2021), pp. 396–405. 10.1007/978-3-030-87589-3_41.