bioengineering

MDPI

Article

Uncertainty-Aware Convolutional Neural Network for Identifying Bilateral Opacities on Chest X-rays: A Tool to Aid Diagnosis of Acute Respiratory Distress Syndrome

Mehak Arora [1,2,*], Carolyn M. Davis [3,4], Niraj R. Gowda [5], Dennis G. Foster [3], Angana Mondal [2], Craig M. Coopersmith [3,4] and Rishikesan Kamaleswaran [2,4,*]

[1] Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
[2] Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA 30332, USA; amonda3@emory.edu
[3] Department of Surgery, Emory University School of Medicine, Atlanta, GA 30332, USA; carolyndavis@emory.edu (C.M.D.); dennis.gene.foster.iii@emory.edu (D.G.F.); cmcoop3@emory.edu (C.M.C.)
[4] Emory Critical Care Center, Emory University School of Medicine, Atlanta, GA 30332, USA
[5] Division of Pulmonary, Critical Care, Allergy and Sleep Medicine, Emory University School of Medicine, Atlanta, GA 30332, USA; niraj.raju.gowda@emory.edu
* Correspondence: marora42@gatech.edu (M.A.); rkamaleswaran@emory.edu (R.K.)

Abstract: Acute Respiratory Distress Syndrome (ARDS) is a severe lung injury with high mortality, primarily characterized by bilateral pulmonary opacities on chest radiographs and hypoxemia. In this work, we trained a convolutional neural network (CNN) model that can reliably identify bilateral opacities on routine chest X-ray images of critically ill patients. We propose this model as a tool to generate predictive alerts for possible ARDS cases, enabling early diagnosis. Our team created a unique dataset of 7800 single-view chest-X-ray images labeled for the presence of bilateral or unilateral pulmonary opacities, or 'equivocal' images, by three blinded clinicians. We used a novel training technique that enables the CNN to explicitly predict the 'equivocal' class using an uncertainty-aware label smoothing loss. We achieved an Area under the Receiver Operating Characteristic Curve (AUROC) of 0.82 (95% CI: 0.80, 0.85), a precision of 0.75 (95% CI: 0.73, 0.78), and a sensitivity of 0.76 (95% CI: 0.73, 0.78) on the internal test set while achieving an (AUROC) of 0.84 (95% CI: 0.81, 0.86), a precision of 0.73 (95% CI: 0.63, 0.69), and a sensitivity of 0.73 (95% CI: 0.70, 0.75) on an external validation set. Further, our results show that this approach improves the model calibration and diagnostic odds ratio of the hypothesized alert tool, making it ideal for clinical decision support systems.

Keywords: chest X-ray classification; computer-aided diagnosis; medical image analysis; model calibration; acute respiratory distress syndrome; deep learning; convolutional neural networks; uncertainty modeling

Citation: Arora, M.; Davis, C.M.; Gowda, N.R.; Foster, D.G.; Mondal, A.; Coopersmith, C.M.; Kamaleswaran, R. Uncertainty-Aware Convolutional Neural Network for Identifying Bilateral Opacities on Chest X-rays: A Tool to Aid Diagnosis of Acute Respiratory Distress Syndrome. *Bioengineering* **2023**, *10*, 946. https://doi.org/10.3390/bioengineering10080946

Academic Editor: Giuseppe Baselli

Received: 30 June 2023
Revised: 26 July 2023
Accepted: 3 August 2023
Published: 8 August 2023

1. Introduction

Acute respiratory distress syndrome (ARDS) is a diffuse lung injury characterized by inflammation leading to increased pulmonary vascular permeability and loss of aerated lung tissue [1]. Clinical hallmarks include hypoxemia and bilateral radiographic opacities [2]. The 2012 Berlin definition of ARDS has been key in supporting clinicians and guiding clinical research [2]. However, morbidity, mortality, and healthcare costs remain unacceptably high, and ARDS remains a diagnostic challenge [3]. Studies show that early recognition of ARDS can help reduce the progression, severity, and potentially lethal clinical sequela [4].

Machine Learning (ML) methods can reliably and robustly learn complex relationships between clinical data, and several research efforts towards using ML for the predictive modeling of medical diseases are underway [5–8]. Similar research efforts for predicting

the early onset of ARDS are ongoing to improve clinical recognition of the syndrome [9–12]. Although there is no single diagnostic test that can rule in or rule out ARDS, one significant limitation both clinically and with ML models is with reading chest radiographs. Interpretation of chest radiographs per the Berlin criteria can be unreliable and subject to inter-rater variability, often leading to missed or delayed ARDS diagnosis [13]. Therefore, we propose a robust and reliable ML model that can be trained to identify bilateral pulmonary opacities on chest X-ray (CXR) images, which could be an invaluable inclusion in the clinical workflow [9]. An automatic system that can raise alerts on detecting lung airspace opacities in chest X-rays of critically ill patients can identify possible ARDS cases, which a physician can then review. It can reduce the burden of reviewing CXR images in situations where there is a high patient-to-clinician ratio and can prevent cases of missed diagnoses that could occur due to human error [14–16].

An important requirement for using data-driven techniques like machine learning to build models for ARDS is a sizeable set of retrospective patient data with known ARDS diagnosis and onset times. Hospital billing codes are insufficient to create such a dataset due to clinical under-recognition [14]. Automated methods to filter patient encounters using constraints on clinical features from the Electronic Medical Record (EMR) like PaO_2/FiO_2 ratio, or positive end-expiratory pressure (PEEP) defined by the Berlin criteria are being approached [17], but are limited by non-standard documentation across EMR [18], and the heterogeneity of clinical manifestations of ARDS [19]. Thus, clinician adjudication of retrospective data is needed, which involves meticulous inspection of patient history using the EMR. An ML model that can flag patients with bilateral pulmonary opacities on chest radiographs as likely candidates for ARDS would make the adjudication process faster with higher ARDS-positive yields. Thus, our proposed tool can also speed up retrospective EMR data adjudication for ARDS cases.

Recently, there has been a surge of interest in using computer vision techniques to identify common findings on CXR images [20–22]. The availability of large CXR datasets such as CheXpert [20] and MIMIC-CXR [21] has made it possible to improve the performance of deep learning models. However, these datasets have labels derived from radiology notes using Natural Language Processing (NLP) Algorithms. These labels can often be inaccurate, owing to variability in radiologists' interpretations, the language used, and differing documentation protocols [23–25]. Label noise can severely impact the true performance of deep learning models [26], and for clinical decision-making tasks, this degradation in performance can come at a dire cost. In this work, we have created a unique dataset of 7800 chest X-ray images with labels generated after careful adjudication of the images for bilateral pulmonary opacities by three blinded, trained physicians.

A review of the recent literature on computer-aided chest X-ray (CXR) classification was performed for the benefit of this study. Numerous studies have adopted convolutional neural networks (CNNs) to identify abnormalities on CXR images, such as pneumonia [27–30], tuberculosis [31–33], and COVID-19 [34–44]. The release of datasets like CheXpert [20] and MIMIC-CXR [21] has also enabled the training of deep learning models on large amounts of data to identify multiple CXR findings like atelectasis, edema, consolidation, cardiomegaly, and pleural effusion. Several high-performing models utilize transfer learning by pretraining CNNs on the ImageNet dataset and fine-tuning them on the target dataset [45]. Notable architectures used for this purpose include ResNets [46,47], DenseNets [20,48,49], Swin Transformers [29], and ConvNeXts [37,50]. Irvin et al. [20] use the Densenet121 architecture to achieve an AUC of 0.90 on the multi-class classification problem. Yuan et al. [48] train models on the CheXpert dataset with the DenseNet121, DenseNet161, DenseNet169, DenseNet201 architectures using a Deep AUC Maximization loss. The DenseNet121 model achieves a superior AUC of 0.93 for a five-class chest X-ray abnormality classification model. Among other notable models, DarkCovidNet [38], based on the DarkNet-19 model used in YOLO object detection, achieves an accuracy of 0.98 for COVID-19 detection. Nahiduzzaman et al. [44] employ a light-weight convolutional neural network coupled with an extreme learning machine model to classify chest X-ray images into 14 classes with an average AUC of 0.96. Yao et al. [51] use

the DenseNet121 as a deep image feature extractor and apply an LSTM (Long Short-Term Memory) network to exploit dependencies between labels, achieving an AUC of 0.79 on a 14-class classification problem using the ChestX-ray14 dataset. Islam et al. [52] use a combined CNN-LSTM framework for COVID-19 detection and achieve an accuracy of 0.99.

Many works incorporate attention mechanisms into CNNs to focus on regions of interest and improve diagnostic accuracy [32,47,49,53]. Many studies use ensembles of machine learning classifiers to improve disease classification using CXR images [29,40,43]. Zhao et al. [49] achieves an AUC of 0.85 on a 14-class chest X-ray classification task, using an attention module added to the DensNet121 architecture and the focal loss for combating class imbalance. Various image processing and augmentation techniques have been used to improve performance and generalization of deep learning models [36,54,55]. Image preprocessing steps include morphological operations for region-of-interest segmentation, machine learning-based lung segmentation, histogram equalization, low-pass filtering, and contrast enhancement [45]. An emerging field of interest involves using generative models to create artificial chest X-ray images for augmenting training datasets. Some researchers have used generative adversarial networks (GANs) for this purpose [56,57], while others have explored multi-modal diffusion models to generate chest X-ray images based on text prompts [58]. These approaches show promise in further enhancing the capabilities of deep learning models for CXR classification.

Research efforts have also been made to specifically diagnose ARDS [10,12,59]. In the work of Reamaroon et al. [10], an image processing-based feature engineering was used in conjunction with deep learning for ARDS detection. Sjolding et al. [12] used a CNN to achieve high performance in identifying ARDS from chest X-rays, and Pai et al. [59] used a multi-modal ensemble framework combining clinical data with chest X-ray imaging to predict ARDS in the first 48 h of admission. A method for dealing with ARDS label uncertainty is proposed by Reamaroon et al. [60], where labels used to train the machine learning model have a confidence score reflecting expert uncertainty in ARDS diagnosis. A similar approach was also followed in the work of Sjolding et al. [12] for ARDS identification. Owing to the complex nature of the syndrome and its diagnosis, we do not attempt to identify ARDS from CXR images in our work but instead identify bilateral pulmonary opacities from CXR images. Previous studies have also worked towards identifying specific lung opacities [23,61]. Vardhan et al. [23] quantify the extent of lung opacities in CXR images, while Kim et al. [61] propose a method to localize opacities in the four quadrants of the lung. In this work, we trained CNN to predict three classes: "bilateral opacities present", "bilateral opacities absent", and "equivocal". Many previous works in CXR image classification disregarded equivocal images or treated them as controls. In contrast, we allow our model to predict the "equivocal" class. We use an uncertainty-aware label-smoothing technique that improves model calibration. Both these training methods ensure that the CNN does not make overconfident predictions on images it is unsure about and can defer these to the physician, making it easy to incorporate in the clinical workflow and potentially improve the rate of ARDS detection [9]. Thus, our model served as a proof of concept of a valuable tool for generating proactive alerts, allowing for the early identification of potential Acute Respiratory Distress Syndrome (ARDS) cases in clinical and research settings.

2. Materials and Methods

2.1. Dataset Generation

This was a single-center retrospective cohort study at an academic institution. We included de-identified patients diagnosed with sepsis based on the Sepsis 3 criteria, admitted to the Surgical and Medical Intensive Care Units (ICU) at Emory University Hospital between August 2015 to May 2019. All patients chosen were over 18 years of age. A total of 7800 single-view frontal chest radiographic images and their corresponding radiologist-dictated reports, corresponding to 664 patient encounters, were extracted from our Electronic Medical Record (EMR) database. These images were annotated by three blinded physicians with critical care experience for pulmonary opacities using the labels: "bilateral",

"left lung", "right lung", or "absent". Unclear images, images with occluded lung regions, or images that could not be interpreted with certainty without further information about a patient's clinical course, were labeled "equivocal". Inter-rater disagreements were resolved conservatively, biased towards limiting false positives. If two annotators agreed, their label would be considered the "ground truth label". If all three clinicians agreed that opacities were present, the image was labeled "bilateral opacities present". If all three clinicians disagreed on the image's label, the image was labeled "equivocal".

2.2. Training the Convolutional Neural Network

2.2.1. Image Preprocessing

We selected all lung-window single-view chest X-ray scans of patient encounters in the chosen cohort and converted them from DICOM format to PNG files. To ensure consistent image analysis, we normalized the image intensity histograms of all chest X-ray images. During the training process of our network, we applied various on-the-fly data augmentation techniques, including random rotations, left-right flips, affine shifts, scaling, and random brightness and contrast adjustments. These augmentation techniques were employed to simulate the inherently noisy nature of chest radiograph images of critically ill patients in the intensive care unit (ICU). Such patients may be unable to maintain a straight posture during the X-ray procedure or may have occluding objects and support devices that cannot be removed. By incorporating these transformations into the training process, we aimed to enhance the robustness of our network to handle real-world variations encountered in ICU CXR scans. During inference and testing, the images were resized to 256×256, histogram normalized, and contrast adjusted using $\gamma = 1.5$.

2.2.2. Network Architecture and Training Parameters

We employed a Convolutional Neural Network (CNN) with the Densenet121 architecture [62], augmented by a fully connected layer, to categorize chest X-ray images into three distinct classes: "bilateral opacities present", "bilateral opacities absent", and "equivocal". We pre-trained the network on the ImageNet dataset to facilitate the training process. Subsequently, we fine-tuned the CNN on the CheXpert dataset, which comprises approximately 191,000 chest X-ray images and associated labels derived from radiology notes. We trained the CNN to identify eight common findings, i.e., lung opacities, edema, consolidation, pneumonia, atelectasis, pneumothorax, pleural effusion, and support devices, using the CheXpert data labels. We further fine-tuned the network on our cohort of 7800 ground-truth annotated images for our specific task, utilizing a patient-level train-test split of 70–30. We used the Adam optimizer [63] with a learning rate of 10^{-4} and trained the CNN for 30 epochs. To address the class imbalance, we utilized a weighted random sampler that combined undersampling of the majority class and upsampling of the minority class, utilizing the class-balanced weights function from the sci-kit-learn Python package. We experimented with different loss functions, including focal loss [64] with a $\gamma = 2$, cross-entropy loss, cross-entropy loss with label smoothing [65], and our proposed cross-entropy loss with uncertainty-aware probability targets. The high-level training diagram of our network is shown in Figure 1. To compare our three-class prediction approach with standard training schemes of predicting only the positive and negative class, we conducted similar experiments by training the CNN to predict only two classes, namely "bilateral opacities present" and "bilateral opacities absent", while disregarding equivocal images as well as while treating them as controls. The outcomes are presented in detail in Section 3 of this report.

We perform additional experiments using different model architectures, as well as ensembling techniques to improve the classification performance of our network [50,66,67]. Details and results about additional experiments we performed to improve diagnostic accuracy are provided in Table S1 of the Supplementary Materials.

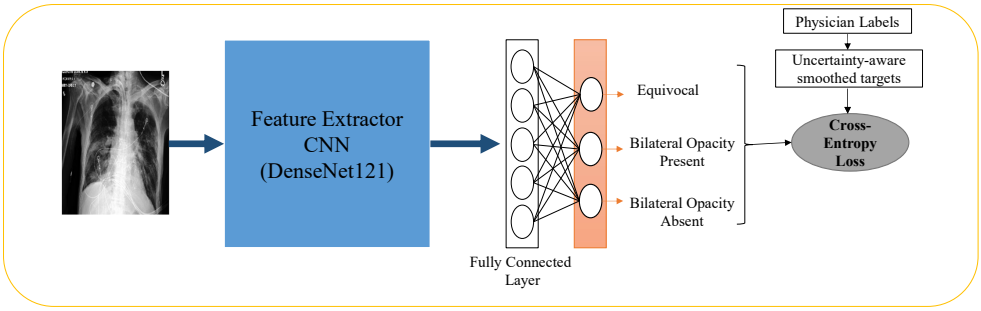

Figure 1. High-level training diagram of the convolution neural network.

2.2.3. Cross-Entropy Loss with Uncertainty Aware Probability Scores

If $p_k = \frac{e^{x^k w_k}}{\sum_{l=1}^{K} e^{x^T w_l}}$ is the likelihood of the k-th class out of K classes, the cross-entropy loss minimizes the function $H(y, p) = \sum_{k=1}^{K} -y_k \log(p_k)$ where y_k are one-hot encoded class labels, with "1" for the correct class and "0" for all other classes. For a network with label smoothing with parameter α, the smoothed cross-entropy loss minimizes the following function $H^s(y^s, p) = \sum_{k=1}^{K} -{y^s}_k \log(p_k)$ where $y^S = y(1-\alpha) + \alpha/K$ [65]. Such soft targets are shown to improve model calibration so that the confidence of predicted probabilities is proportional to the likelihood of the prediction being accurate [65]. For our three-class model predicting bilateral opacities present, absent, and equivocal images, we tweak the soft targets to reflect the probabilities of the respective classes. In essence, if the ground truth label is "equivocal", the target likelihoods of the "present" and "absent" classes are set to 0.5. Labels are one-hot encoded by assigning a (1, 0, 0) target label if bilateral opacities are absent, a (0, 1, 0) target label if bilateral opacities are present, and a (0.5, 0.5, 1) target label if the image is marked equivocal. Our results show that this training method allows the CNN to predict the equivocal class for uncertain input images, reducing the rate of false positives and false negatives while maintaining good sensitivity and precision. The CNN trained with uncertainty-aware probability scores performed the best and was chosen for further analysis and external validation.

2.3. Performance Metrics

We conducted a comparative analysis of our convolutional neural network models based on standard performance metrics. These metrics include the Area under the Receiver Operating Characteristics Curve (AUROC) and the Precision–Recall Curve (AUPRC), balanced accuracy score, precision, sensitivity, specificity, F-score, and the Diagnostic Odds Ratio (DOR). These are defined as follows:

$$\text{precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{1}$$

$$\text{sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{2}$$

$$\text{specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \tag{3}$$

$$\text{balanced accuracy} = \frac{\text{sensitivity} + \text{specificity}}{2} \tag{4}$$

$$\text{DOR} = \frac{\text{sensitivity} \times \text{specificity}}{(1 - \text{sensitivity}) \times (1 - \text{specificity})} \tag{5}$$

These metrics capture a more comprehensive view of model performance and utility in the clinical workflow. The diagnostic odds ratio ranges from zero to infinity, and higher

diagnostic odds ratios indicate better test performance [68]. The balanced accuracy score is used as it is a better estimate of model performance in the case of multi-class classification with class imbalance.

2.4. Model Calibration

In assessing the test performance of our models, we considered model calibration as a crucial criterion. Calibration refers to how well the predicted likelihoods from the model align with the actual confidence of the predictions being accurate. To quantify the calibration, we utilized confidence binning and calculated the Maximum Calibration Error (MCE). To apply confidence binning, the predicted likelihoods generated by the model are divided into ten bins at regular. For each bin, we computed the average prediction likelihood and accuracy of the predictions falling within that bin. The MCE, the maximum difference between the average predicted likelihood and the average accuracy across all bins was chosen as the calibration measure. This metric indicates the worst-case calibration error, making it particularly relevant for clinical decision-making tasks where reliable confidence measures are essential.

2.5. Model Interpretability

We employed two interpretability techniques for deep image classification, namely Grad-CAM [69] (Gradient-weighted Class Activation Mapping) and Occlusion Sensitivity Maps [70], to understand the classification decisions made by our CNN model. Grad-CAM is a visualization method that generates heatmaps, indicating the regions of an image most influential in the CNN's classification decision. Gradients of the target class score with respect to the feature maps from the final convolutional layer of the CNN are calculated and used to assign weights to the feature maps, resulting in a heatmap highlighting the discriminatory regions. Occlusion Sensitivity Maps involve gradually occluding different parts of an image and observing the resulting changes in the CNN's output. This technique helps identify the regions within the image that significantly contribute to the CNN's classification decision and regions that decrease the CNN prediction confidence.

2.6. External Dataset Validation

We use the MIMIC-CXR dataset [21] for external validation of our model. To do so, we first matched chest X-ray images and radiology notes to all patients from the MIMIC-IV dataset that were diagnosed with sepsis-3. We then randomly sampled approximately 1600 images from this subset and obtained physician labels using the same strategy described in Section 2.1. These are considered ground truth labels for validating our algorithm on the external dataset. The chest X-ray scans undergo the same pre-processing steps and test time transformations as those applied to our internal testing dataset. These include resizing the image to a 256×256, histogram normalization, and contrast adjustment to a $\gamma = 1.5$.

2.7. Comparison with Labels Derived from Radiology Notes

Our study used a rule-based Natural Language Processing (NLP) pipeline to analyze the radiology notes associated with each chest X-ray scan. We followed the methodology outlined by Irvin et al. [20] and designed an NLP pipeline using the medspaCy [71] library, a framework specifically tailored for clinical text processing. The medspaCy pipeline incorporated various components to facilitate the analysis. Firstly, we employed the in-built sentencizer module for text preprocessing, enabling the segmentation of the radiology notes into individual sentences. We then performed Named Entity Recognition using a target matcher component to identify specific entities of interest within the text. Additionally, we integrated a context module into the pipeline to capture the contextual information related to the identified named entities, such as negation or likelihood. To structure the analysis, we included a sectionizer component as a pre-processing step to demarcate different sections within the radiology notes, including clinical indications, impressions, and findings pertaining to the lungs/pleura, heart/mediastinum, and bones/soft tissue.

Additional context and target rules derived from the language syntax frequently seen in our cohort of radiology notes were incorporated, as were phrases and context rules described by the Stanford CheXpert Labeler team. For the identification of possible cases of pulmonary opacification, we employed specific keywords, including ARDS, atelectasis, consolidation, edema, effusion, opacity, infiltrates, pneumonia, and pneumothorax. These keywords were used to flag instances in the radiology notes that potentially indicated the presence of lung opacities, and the associated images were subsequently filtered out based on mentions of these findings in associated radiology notes. They were then categorized into positive (indicating the presence of the finding), negative (indicating the absence of the finding), or uncertain (suggesting the possibility of existence, requiring further clinical adjudication). Furthermore, we detected disease descriptors, such as 'bilateral' for 'opacity', 'infiltrates', and 'pleural' or 'interstitial' for 'effusions' or 'edema', respectively. To label bilateral opacities, we utilized descriptors of laterality ('Bilateral', 'Right', 'Left'). Instances describing findings as only 'Right' or 'Left' were labeled negative for bilateral opacities. We compared the performance of the labels derived from these radiology notes with the output of our CNN model, which was trained on physician labels. The obtained results are discussed in detail in Section 4 of our study.

3. Results

3.1. Data Characterization

The number of patients, chest radiographs, image label frequency, and demographic information of the internal and external datasets are listed in Table 1. Class labels in both datasets followed a long-tailed distribution, with the lowest prevalence of the equivocal class. Since our patient cohort were all those diagnosed with sepsis, we can see a higher incidence of the bilateral pulmonary opacities present class when compared to the absent and equivocal classes, with 54% present in the internal (Emory University) dataset and 77% present in the external (MIMIC-CXR) dataset.

Table 1. Demographic information of the internal test set from Emory University and the external validation set (MIMIC-CXR). Data are reported as the number of counts for patients and CXR images, number of counts (percentage of total counts) for sex, race, and label prevalence, or median (IQR) for Age.

		Internal Dataset (Emory University)	**External Dataset (MIMIC-CXR)**
Patients		663	952
Chest X-rays		7825	1639
Age—median (IQR)		57 (43–67)	65 (53–76)
Sex	Male	315 (48%)	532 (56%)
	Female	348 (52%)	420 (44%)
Race	Caucasian/White	304 (46%)	605 (63%)
	African American/Black	251 (38%)	153 (16%)
	Asian	14 (2%)	44 (5%)
	Other	94 (14%)	150 (16%)
CXR Labels	Bilateral Opacities Present	4227 (54%)	1009 (61.5%)
	Bilateral Opacities Absent	1788 (23%)	442 (27%)
	Equivocal	1810 (23%)	188 (11.5%)

3.2. CNN Performance Comparison

Table 2 reports the performance metrics of the evaluated training schemes. All the CNN models compared are those pre-trained on the ChexPert dataset and then had all

layers fine-tuned on the Emory Dataset. These models saw a superior performance to those trained from scratch, those pre-trained on the CheXpert dataset, and those with their last layers fine-tuned on the Emory dataset. The metrics in Table 2 and the Receiver Operating Characteristics (ROC) curves and Precision–Recall Curves (PRC) in Figure 2 are calculated for the "Bilateral Opacities Present" class to enable comparison between the two-class and the three-class approach. The three-class model refers to the training scheme in which we train the model to predict the "present" or "absent" as well as the "equivocal" class. The two-class model refers to the training scheme in which we predict only the "present" and "absent" classes. Images labeled "equivocal" as belonging to the "absent" class in one set of experiments were disregarded and removed from analysis for a second set.

The results in Table 2 show that the highest performance metrics were observed when testing the two-class model (disregarding equivocal). However, this might represent a biased testing set as we eliminated the difficult examples from the test set. Thus, this approach was not considered for performance comparison or external validation. The three-class model outperforms the two-class model (with equivocal images treated as controls) in precision, AUPRC, and diagnostic odds ratio. Superior performance was achieved by training the three-class model using the cross-entropy loss with uncertainty-aware probability targets, with an AUROC of 0.828 (95% CI: 0.803, 0.853) and F-score of 0.746 (95% CI: 0.720, 0.775), and a diagnostic odds ratio of 17.211 (95% CI: 12.106, 25.976). The Receiver Operating Characteristics curves (ROC) and Precision–Recall Curves (PRC) in Figure 2 highlight the benefit of our three-class approach in improving the precision and recall of the trained model. The benchmark performance of labels derived from radiology notes showed an average precision of 0.63, an average sensitivity of 0.58, and an average specificity of 0.81. Results of a 10-fold cross-validation are provided in Table S2 of the Supplementary Materials.

Table 2. Performance metrics of the two-class model (predicting "present" or "absent") with equivocal images considered "absent", two-class model (disregarding equivocal) with equivocal images disregarded from training and testing, and the three-class model (predicting "present", "absent" and "equivocal"). The highest values for each metric are emboldened, excluding the two-class model (disregarding equivocal) to ensure valid and unbiased comparisons. A 95% CI is reported and was calculated by bootstrap resampling 1000 times.

Experiment	Loss Function	AUROC	AUPRC	F-Score	Precision	Sensitivity	Specificity	Diagnostic Odds Ratio	Balanced Accuracy
Two-Class Model	Cross-Entropy Loss	0.819 (0.791, 0.844)	0.553 (0.521, 0.584)	0.707 (0.679, 0.735	0.678 (0.671, 0.726)	0.724 (0.694, 0.754)	0.809 (0.781, 0.836)	11.141 (8.545, 14.765)	0.767 (0.736, 0.795)
	Focal Loss	0.825 (0.802, 0.852)	0.568 (0.539, 0.595)	0.715 (0.684, 0.744)	0.706 (0.674, 0.735)	0.730 (0.698, 0.732)	0.816 (0.787, 0.843)	11.967 (8.979, 16.144)	0.773 (0.743, 0.788)
Two-Class Model^{DE}	Cross-Entropy Loss	0.884 (0.864, 0.904)	0.758 (0.736, 0.779)	0.734 (0.689, 0.743)	0.777 (0.750, 0.807)	0.714 (0.689, 0.743)	0.923 (0.904, 0.941)	29.917 (22.455, 43.151)	0.819 (0.796, 0.842)
	Focal Loss	0.875 (0.855, 0.896)	0.756 (0.731, 0.789)	0.774 (0.746, 0.799)	0.775 (0.750, 0.803)	0.772 (0.742, 0.801)	0.868 (0.843, 0.893)	22.235 (16.398, 31.375)	0.820 (0.793, 0.847)
Three-Class Model	Focal Loss	0.810 (0.785, 0.837)	0.845 (0.803, 0.8475)	0.739 (0.712, 0.766)	0.738 (0.711, 0.766)	0.740 (0.712, 0.767)	0.712 (0.666, 0.757)	7.006 (5.011, 10.191)	0.726 (0.689, 0.762)
	Cross-Entropy Loss	0.790 (0.764, 0.818)	0.819 (0.0.792, 0.834)	0.688 (0.659, 0.717)	0.716 (0.691, 0.744)	0.707 (0.681, 0.732)	**0.851 (0.819, 0.883)**	13.769 (9.948, 19.926)	0.776 (0.750, 0.808)
	Cross-Entropy Loss with probability targets	**0.828 (0.803, 0.853)**	**0.874 (0.861, 0.887)**	**0.746 (0.720, 0.775)**	**0.755 (0.731, 0.782)**	**0.761 (0.735, 0.786)**	0.842 (0.812, 0.875)	**17.211 (12.106, 25.976)**	**0.802 (0.773, 0.831)**

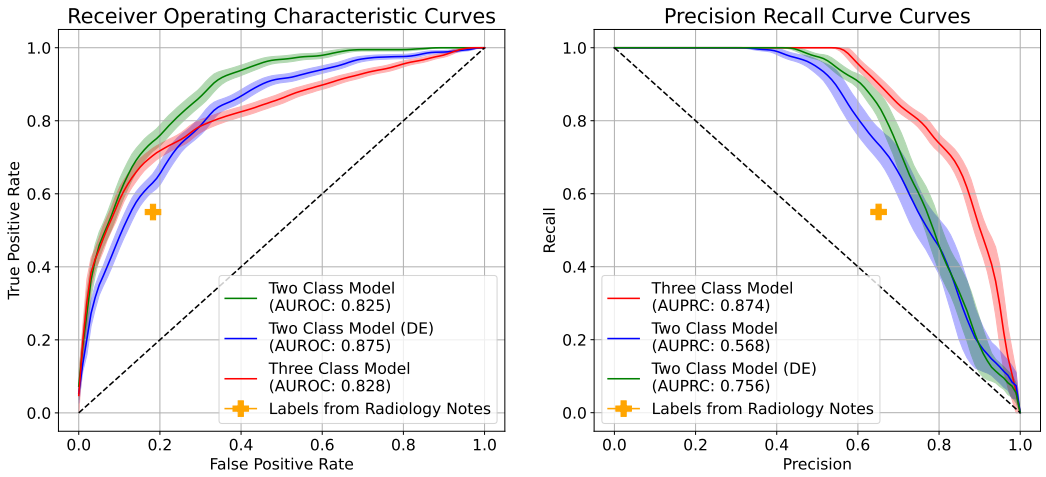

Figure 2. Receiver Operating Characteristics curves (ROC) and Precision–Recall Curves (PRC) for the positive class (bilateral opacities present) of the two-class and three-class models, as well as the benchmarked performance of the labels derived from radiology notes using an NLP algorithm. DE: Disregarded Equivocal images.

3.3. Model Calibration

Table 3 reports the maximum calibration error for the experiments described in the previous section. Figure 3 plots reliability diagrams that depict model calibration. It is observed that the three-class model trained using the cross-entropy loss with probability targets has the lowest MCE of 0.150 with the best-calibrated reliability diagram. The three-class model trained with focal loss is a close second, with an MCE of 0.240. It can be noticed that the two-class models have inferior model calibration when compatible with the three-class models.

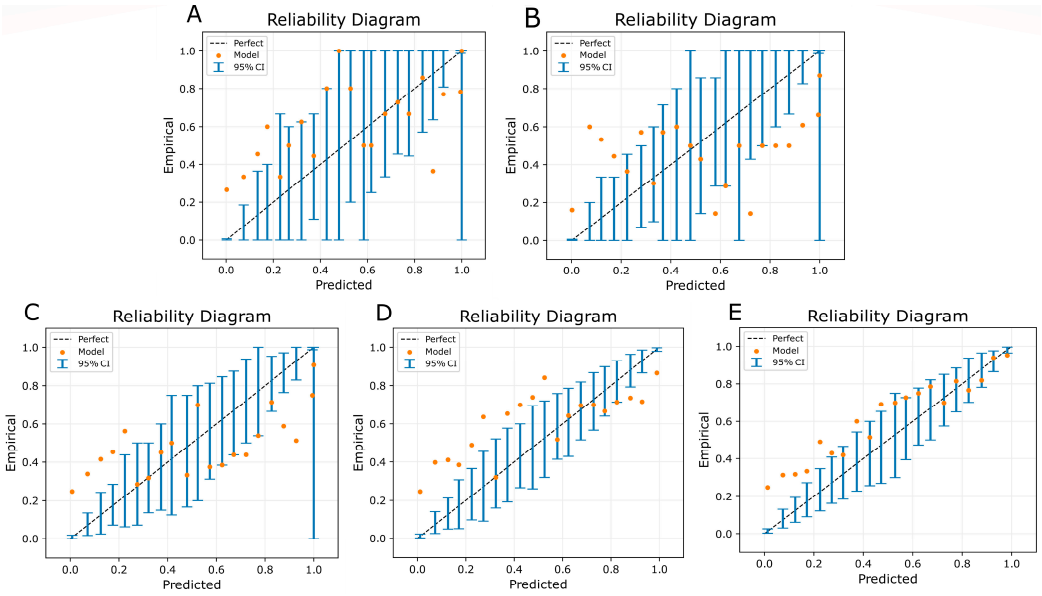

Figure 3. Reliability diagrams to inspect model calibration for (**A**) two-class model trained with cross-entropy loss, (**B**) two-class model trained with focal loss, (**C**) three-class model trained with cross-entropy loss, (**D**) three-class model trained with focal Loss, (**E**) three-class model trained with cross-entropy loss using uncertainty-aware probability targets.

Table 3. Maximum calibration errors of the two-class model (predicting "present" or "absent") with equivocal images considered "absent", two-class model (disregarding equivocal) with equivocal images disregarded from training and testing, and the three-class model (predicting "present", "absent" and "equivocal").

Experiment	Loss Function	MCE
Two-Class Model	Cross-Entropy Loss	0.424
	Focal Loss	0.385
Two-Class Model: Disregarding Equivocal	Cross-Entropy Loss	0.479
	Focal Loss	0.318
Three-Class Model	Cross-Entropy Loss	0.408
	Focal Loss	0.240
	Cross-Entropy Loss with probability targets	**0.150**

3.4. External Validation on MIMIC-CXR

The three-class model trained with the cross-entropy loss using uncertainty-aware probability targets was chosen as the best-performing model and used for external validation. Table 4 reports the performance metrics of the model on the external validation set, while Figure 4 plots the Receiver Operating Characteristics curves and Precision–Recall Curves for the same. The model achieves an AUROC of 0.836 (95% CI: 0.811, 0.858), an F-score of 0.658 (95% CI: 0.627, 0.685), and a diagnostic odds ratio of 2.18 (95% CI: 1, 1.8). Table 1 shows us that the external validation set used for our study has a much higher incidence of the "Bilateral Opacities Present" class. This could explain the reason behind observing a higher AUC and specificity but a lower sensitivity and diagnostic odds ratio. The benchmark performance of labels derived from radiology notes showed an average precision of 0.66, an average sensitivity of 0.59, and an average specificity of 0.58.

Table 4. Performance Metrics for the three-class model trained with cross-entropy loss using uncertainty-aware probability targets on the MIMIC-CXR external validation set. A 95% CI is reported and was calculated by bootstrap resampling 1000 times.

Experiment	AUROC	AUPRC	F-Score	Precision	Sensitivity	Specificity	Diagnostic Odds Ratio	Balanced Accuracy
Three-Class Model: Cross-Entropy Loss with Probability Targets	0.834 (0.811, 0.858)	0.898 (0.873, 0.917)	0.658 (0.627, 0.685)	0.729 (0.704, 0.749)	0.727 (0.703, 0.747)	0.955 (0.929, 0.973)	2.180 (1.801, 1.00)	0.841 (0.816, 0.860)

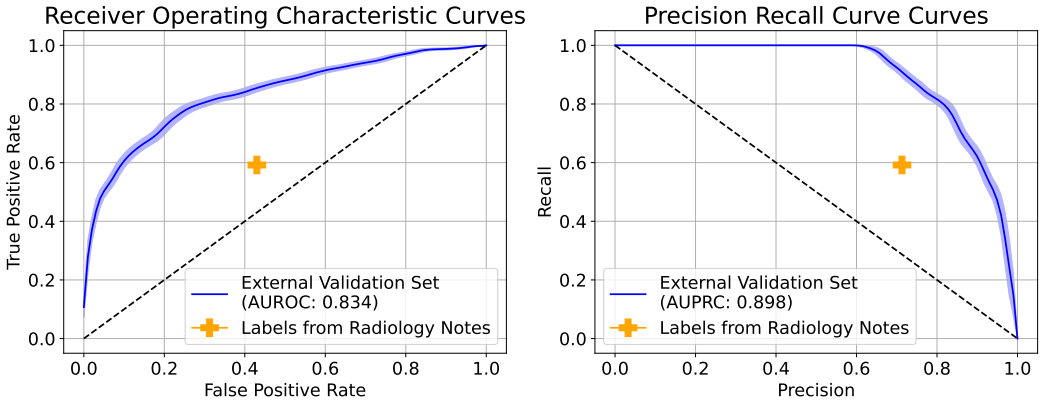

Figure 4. Receiver Operating Characteristics curves (ROC) and Precision–Recall Curves (PRC) for the positive class (bilateral opacities present) of three-class Model trained with cross-entropy loss using uncertainty-aware probability targets on the MIMIC-CXR external validation set, as well as the benchmarked performance of the labels derived from radiology notes using an NLP algorithm.

3.5. Visualization of Saliency Maps

By utilizing GradCAM and occlusion sensitivity maps, we aimed to gain insights into the decision-making process of our CNN model and understand the specific regions of the input images that were most influential in determining the classifications. These visualization techniques enhance our understanding of the model's reasoning and aid in interpreting its predictions. In Figure 5, we can see that while the Grad-CAM either focuses on the entire lung region or each lung separately, the occlusion sensitivity maps show us the exact regions of the image that were important when making the positive prediction. These regions correspond well with opacified lung spaces. In Figure 5a, an interesting observation is that the Grad-CAM clearly pays attention to both lungs and correctly classifies the image as negative for bilateral opacities, despite one lung looking opacified. In image sets (c) and (f), the grad-CAM activations focus on lung regions that are occluded or difficult to interpret with certainty due to poor quality. Figure 6 displays instances of misclassification compared to the gold standard of physician labels. Image sets (c) and (d) are examples of the CNN classifying images that are labeled equivocal as "Bilateral Opacities Absent" and "Bilateral Opacities Present", respectively, albeit with lower likelihoods. Image sets (e) and (f) are examples of the CNN classifying certain images that it is uncertain about as "equivocal". We notice that the Grad-CAM appears to be focusing on the breast tissue region, which might confuse the predictions of airspace opacities.

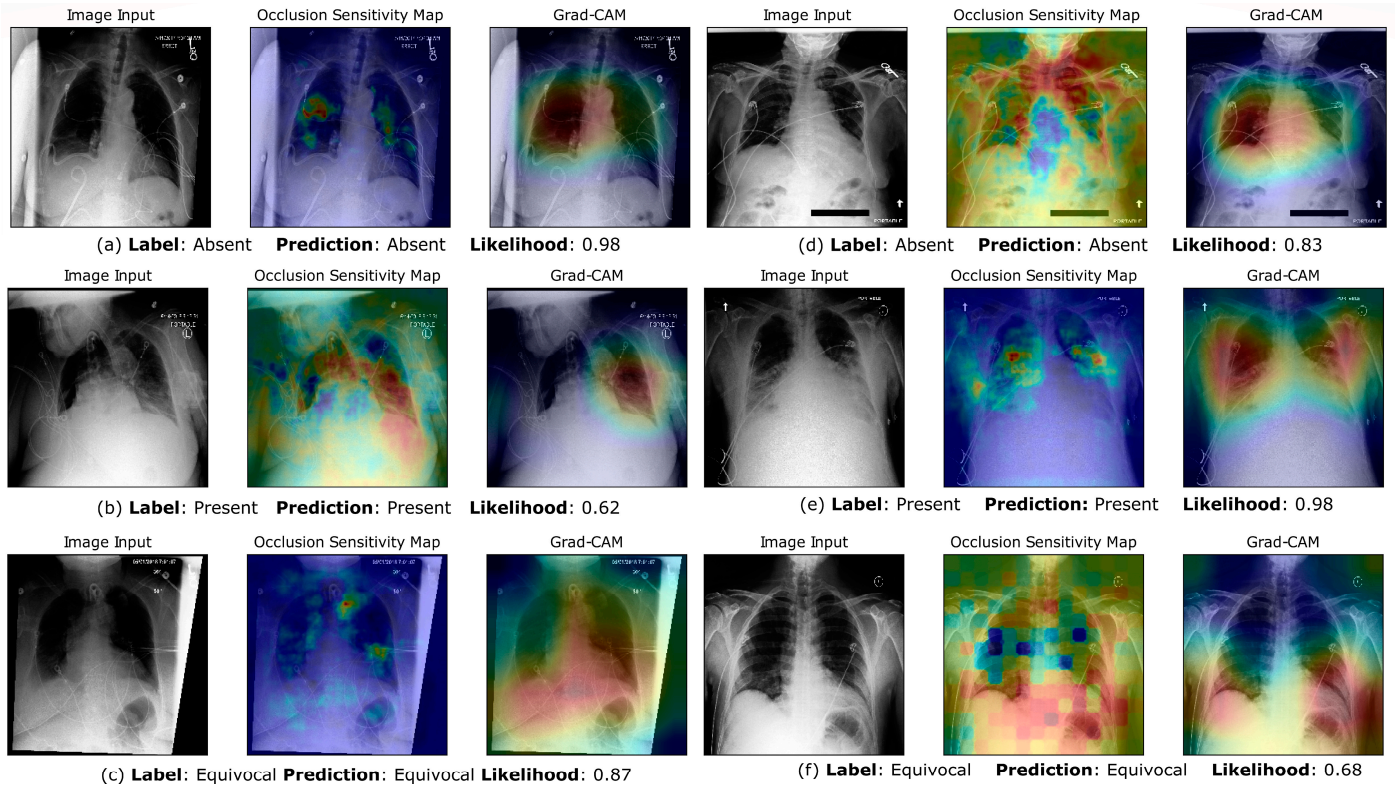

Figure 5. Occlusion Sensitivity Maps and Grad-CAM visualizations of network predictions for correctly classified examples. (**a**–**c**) are CXR images belonging to the internal test set, while (**d**–**f**) are CXR images belonging to the external validation set. For the Grad-CAM visualizations, the red highlighted regions have the highest discriminatory importance for the predicted class. For the Occlusion Sensitivity Maps, the hotter (red) regions are the areas that increased model confidence in the predicted class. In contrast, the cooler (blue) regions were the most confusing to the CNN (increased predicted class likelihood when occluded).

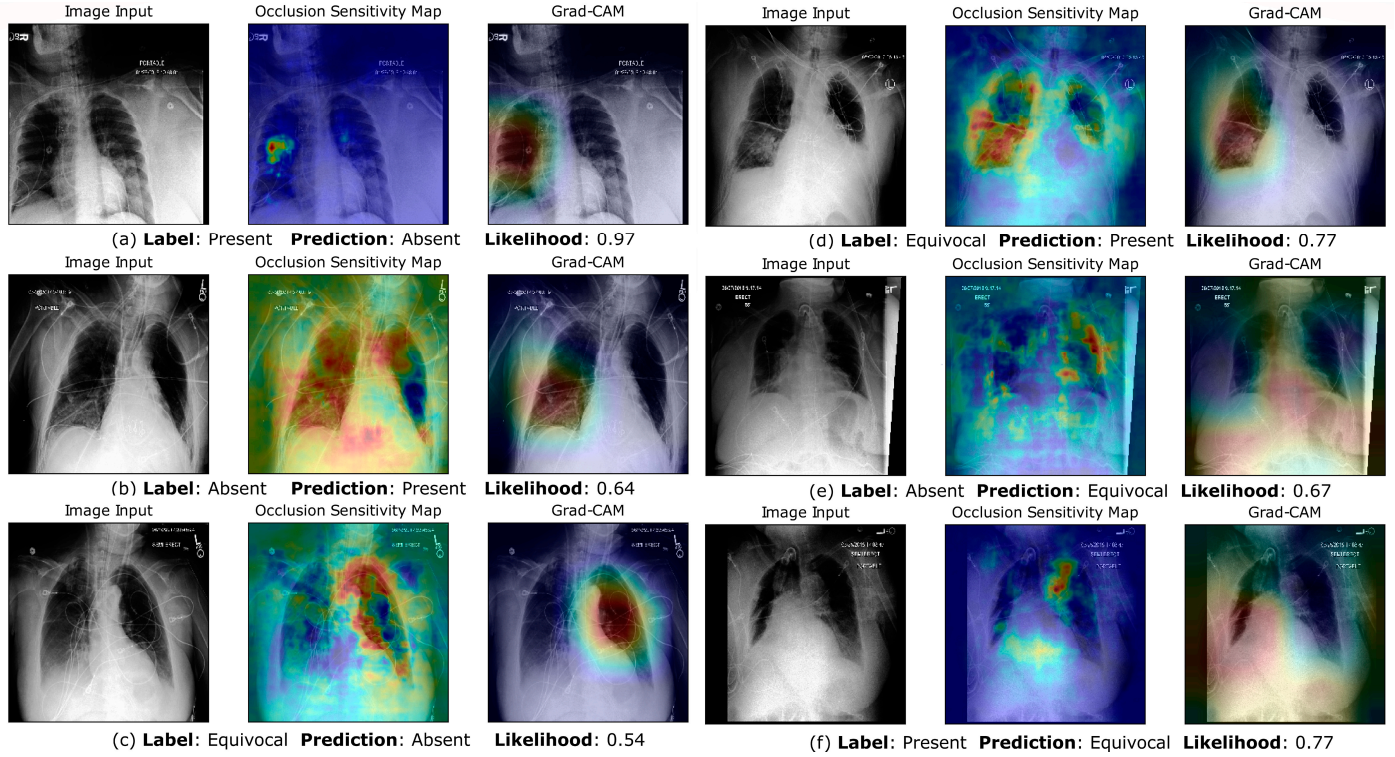

Figure 6. Occlusion Sensitivity Maps and Grad-CAM visualizations of network predictions for misclassified examples. (**a**,**b**) Examples of misclassification between the "Bilateral Opacities Present" and "Bilateral Opacities Absent" Class. (**c**,**d**) Examples of CXR images labeled "equivocal" by clinicians, that the network identifies as "Bilateral Opacities Present" and "Bilateral Opacities Absent". (**e**,**f**) Examples of CXR images labeled "Bilateral Opacities Present" and "Bilateral Opacities Absent", that the network identifies as "equivocal".

4. Discussion

Recent research on ARDS has moved towards early recognition as a means of reducing severity and mortality, and the initiation of therapies known to reduce the likelihood of ARDS development [72], like low tidal volume ventilation [73] and fluid restriction [74]. Critically ill patients have frequent chest X-rays taken to ensure the correct placement of devices such as chest tubes, central lines, or pacemakers, and for routine medical workups [75]. Using these images to help with the accurate and early identification of bilateral pulmonary opacities can optimize patient care and ongoing clinical management, possibly improving the chances of detecting ARDS in the early stages [9]. In this work, we have proposed a CNN trained to identify pulmonary opacities on chest X-ray images as a tool to generate predictive alerts for possible ARDS cases that can be applied in the clinical and research domains. The overarching goal of our research was to develop a robust and dependable model capable of detecting bilateral opacities on chest radiographs to translate it into clinical settings. The clinical utility of our work is threefold.

First, we have collected a dataset comprising all chest X-ray (CXR) images of patients diagnosed with sepsis-3 during their hospitalization at Emory University Hospital. We did not curate the dataset by selectively choosing good-quality images or by sampling images from specific 'present' and 'absent' classes. This approach stands in contrast to some recent studies focusing on disease detection, in which datasets are composed of positive disease samples, and controls [34,35,38,52]. Some studies struggle with poor generalization when applied in real clinical settings [76,77]. The lack of representation of critical patients with possible differential diagnoses and the differences in inclusion criteria for datasets can be confounding factors that contribute to this disparity in performance. Instead, we aimed to maintain a reflection of real-world data distributions by including all available images.

Consequently, our patient cohort represents a high-acuity group, and a significant portion of the images received equivocal labels as determined by the clinical adjudicators (23%). Images in our data are often of poor quality, as they are often captured at the bedside in the ICU, with support devices, chest tubes, and other occluding objects.

Second, we generate ground truth labels for all images from three blinded clinicians to train the network instead of using labels extracted from radiology notes as these might not always be accurate or reflective of a patient's current clinical status [15,24,25]. Findings that are repetitive or beyond the indication, such as line placement confirmation, often go unlabeled completely. In the case of labeling for bilateral opacities, the problem becomes more challenging owing to variability in the interpretation of the criterion. Additionally, radiologists may omit mentions of laterality or might use non-standard language to indicate findings. We have also observed similar evidence of unreliable labels from radiology notes, which can be seen in Figure 7. In this work, we create a first-of-its-kind dataset of single-view chest X-ray images of critically ill patients annotated for bilateral opacities, unilateral opacities, or labeled equivocal by three clinical experts accounting for inter-rater variability. Since the presence of bilateral opacities on chest radiographs is one of the primary criteria of the Berlin Definition for ARDS, such a dataset could be pivotal in predictive modeling for ARDS using machine learning algorithms. The rigorous process of clinician adjudication guarantees that the performance metrics reported in our study hold clinical significance and can be trusted in real-world medical applications. By relying on ground truth labels established through consensus among multiple experts, we provide a reliable foundation for training our network and validating its performance.

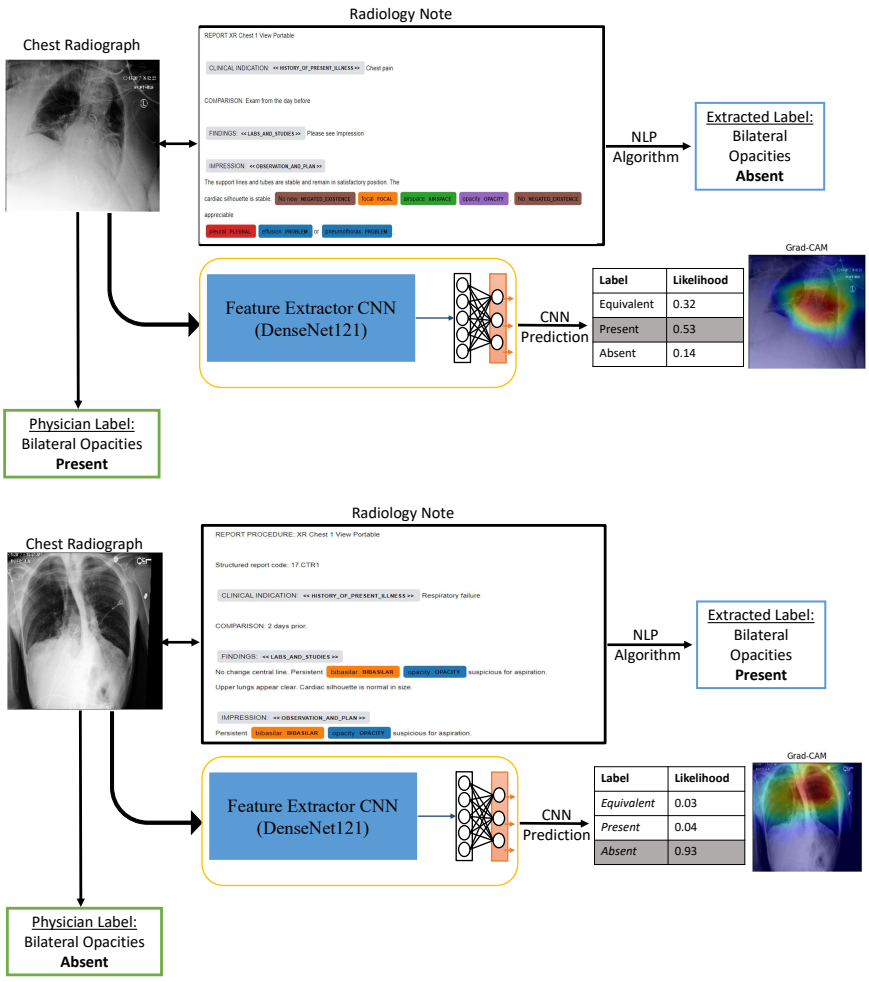

Figure 7. Illustrative examples of incorrect labels derived from radiology notes.

Third, our machine learning algorithm attempts to incorporate trust and uncertainty awareness by modeling the "equivocal" class. We allow clinicians to label images as equivocal and design a three-class training scheme incorporating label uncertainty and explicitly teaching CNN to identify the equivocal class. This approach was adopted to develop a reliable model that can label an image as 'equivocal' rather than producing low-confidence predictions on uncertain cases. By training our model to handle equivocal labels, we aimed to create a system that can be seamlessly integrated into the clinical workflow or employed in a human-in-the-loop setting. Unlike many current state-of-the-art studies, which focus on improving the AUROC, our training goal was to produce a well-calibrated model, incorporating uncertainty awareness via the "Equivocal" class. The output probabilities of a well-calibrated model are good estimates of its predictive confidence. This ensures that the model's predictions are reliable and can be interpreted with greater confidence in real-world clinical applications. To demonstrate the superiority of our three-class model over a two-class model that treats equivocal labels as "bilateral opacities absent", we conducted an experimental comparison. The three-class model showed higher AUROC, Area Under the Precision–Recall Curve (AUPRC), F-score, calibration error, and diagnostic odds ratio compared to the two-class model. This indicates that our approach of explicitly handling equivocal cases leads to better performance in various evaluation metrics. Moreover, we evaluated different loss functions and found that our uncertainty-aware cross-entropy loss with probability targets achieved a lower maximum calibration error when compared to focal loss and standard cross-entropy loss. This is evident from our results in Table 3 and Figure 3, further validating the effectiveness of our training scheme. The diagnostic odds ratio balances a trade-off between sensitivity and specificity, thereby balancing the trade-off between false positive rates (probability of false alarms) and false negative rates (miss rates). This measure helps evaluate the diagnostic accuracy of the models. Our results show that the three-class model has higher precision, sensitivity, AUPRC, and diagnostic odds ratio when compared to the corresponding two-class models (Table 2).

On-par classification performance on the external validation dataset reiterated the robustness of our approach. We noticed an increase in validation performance in terms of the AUC, which could be an artifact of the much higher number of positive class cases in the external test set. Grad-CAM and occlusion sensitivity map visualizations were used to lend interpretability to class predictions. Figures 5 and 6 show us that the CNN is picking up on the regions of the lungs that are important in identifying bilateral opacities. This confirms that the CNN is trained to look at the relevant areas of the image instead of finding "shortcuts" by picking up accidental statistical correlations with irrelevant regions in the images, which is a growing concern with large datasets from a limited number of hospital centers [77]. Similar saliency maps for the external dataset prove the reliability and robustness of our model.

Our study has some limitations that need to be addressed. First, we limited our dataset to chest X-ray images from a cohort of patients diagnosed with sepsis. This was done to prevent the inclusion of trauma cases, and other causes of ARDS, whose chest X-rays might look widely varied. In future work, we plan to include all critically ill patients admitted to the ICU to train our models on a more diverse patient population. Second, we obtain gold-standard ground truth labels on the external MIMIC-CXR dataset in the same method as the internal validation set due to the lack of pre-available ground truth labels for "bilateral" chest X-ray opacities. Such an external validation might not account for inherent bias in our labeling scheme. However, since three blinded clinicians provide our labels, we consider this an unbiased physician-annotated X-ray read. Another limitation of work might be that our dataset is annotated by physicians instead of radiologists. However, all three physicians have critical care experience. In clinical practice, images obtained at the bedside are often interpreted in real time and factor in the patient's current clinical status, which is often not reflected in the electronic medical record in real-time. Additionally, collaborative discussions between critical care providers and radiologists are not reflected in radiographic annotations. Taken together, the advantage of training our model using

annotated images by three critical care-trained physicians is more reflective of actual clinical practice, potentially improving the translatability of our model into clinical practice.

Moreover, this work is a proof-of-principle of an automated tool that can be used to check the chest radiograph criteria of the Berlin definition of ARDS. However, ARDS is a syndrome that needs a multi-modal approach for diagnosis. As a next step, we aim to look at a series of chest X-ray scans of a particular patient and integrate clinical vitals, labs, and blood gas readings to build a better predictive model for ARDS diagnosis.

In future work, the inclusion criteria for our study will be explained to include CXR images of all patients admitted to the Emory University Hospital from 2016–2022, to enhance accuracy in our model. We will explore using generative models to augment the training data by producing artificially generated CXR image samples. Additional considerations to this approach include acquiring clinician-validated "ground-truth" labels for artificially generated images. Attention-based convolutional neural networks will be considered, in addition to transformer-based methods to improve performance.

5. Conclusions

Acute Respiratory Distress Syndrome (ARDS) is a serious lung injury with high mortality that is often missed or diagnosed late, leading to poor patient outcomes. An important criterion for diagnosing ARDS is the presence of bilateral pulmonary opacification. In this work, we created a unique dataset of single-view CXR images of critically ill patients labeled as bilateral opacities, unilateral opacities, and equivocal by three clinical experts with inter-rater agreement. Our model is robust to clinically equivocal images and explicitly flags images as equivocal, indicating the need for more information or physician input. Our training approach using uncertainty-aware soft likelihood targets provides reliable output probabilities calibrated to confidence in the presence of bilateral pulmonary opacities. The CNN model identifies bilateral opacities with high precision and a low false-positive and false-negative rate. Thus, our model design is well-suited for an automated, always-on alert system that can tirelessly read and make predictions on chest X-rays with high confidence while deferring equivocal images to clinicians on-call. Such an AI-clinician collaborative strategy for ARDS onset prediction can be a useful tool for early identification and for retrospective adjudication for data-driven studies.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/bioengineering10080946/s1, Table S1: Performance Metrics of the Three Class Model (predicting "present", "absent" and "equivocal"). The highest values for each metric are emboldened. 95% Confidence Intervals are reported and were calculated using non-parametric estimation using bootstrap-ping; Table S2. Results of 10-fold cross-validation using the DenseNet121 model trained with uncertainty-aware cross-entropy loss with probability targets. References [42,45,50,66,67] are cited in the Supplementary Materials.

Author Contributions: Conceptualization, R.K., M.A., C.M.D.; methodology, R.K., M.A., C.M.D.; implementation and investigation, M.A., A.M.; formal analysis, M.A., C.M.D., R.K.; data curation, C.M.D., N.R.G., D.G.F.; supervision, R.K., C.M.C.; writing—original draft preparation, M.A.; writing—review and editing, C.M.D., R.K., C.M.C., A.M., N.R.G.; visualization, M.A.; funding acquisition, R.K., C.M.C. All authors have read and agreed to the published version of the manuscript.

Funding: R.K. and M.A. were supported by the National Institutes of Health (NIH) Award Number R01GM139967 and UL1TR002378. C.M.D. and C.M.C. were supported by NIH Award Number GM148217 and GM095442.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Emory University (IRB# STUDY00000302).

Informed Consent Statement: Not applicable.

Data Availability Statement: Implementation code for this work is available at https://github.com/Kamaleswaran-Lab/CXaRds, accessed 1 August 2023. Data can be available by contacting the authors.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ARDS	Acute Respiratory Distress Syndrome
CNN	Convolutional Neural Network
ML	Machine Learning
EMR	Electronic Medical Record
CXR	Chest X-rays

References

1. Ware, L.B.; Matthay, M.A. The acute respiratory distress syndrome. *N. Engl. J. Med.* **2000**, *342*, 1334–1349. [CrossRef] [PubMed]
2. The ARDS Definition Task Force. Acute Respiratory Distress Syndrome. *JAMA* **2012**, *307*, 2526–2533. [CrossRef]
3. Bellani, G.; Laffey, J.G.; Pham, T.; Fan, E.; Brochard, L.; Esteban, A.; Gattinoni, L.; van Haren, F.; Larsson, A.; McAuley, D.F.; et al. Epidemiology, Patterns of Care, and Mortality for Patients With Acute Respiratory Distress Syndrome in Intensive Care Units in 50 Countries. *JAMA* **2016**, *315*, 788–800. [CrossRef]
4. Kerchberger, V.E.; Brown, R.M.; Semler, M.W.; Zhao, Z.; Koyama, T.; Janz, D.R.; Bastarache, J.A.; Ware, L.B. Impact of Clinician Recognition of Acute Respiratory Distress Syndrome on Evidenced-Based Interventions in the Medical ICU. *Crit. Care Explor.* **2021**, *3*, e0457. [CrossRef] [PubMed]
5. Zhou, A.; Raheem, B.; Kamaleswaran, R. OnAI-Comp: An Online AI Experts Competing Framework for Early Sepsis Detection. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2021**, *19*, 3595–3603. [CrossRef] [PubMed]
6. van Wyk, F.; Khojandi, A.; Mohammed, A.; Begoli, E.; Davis, R.L.; Kamaleswaran, R. A minimal set of physiomarkers in continuous high frequency data streams predict adult sepsis onset earlier. *Int. J. Med. Inf.* **2019**, *122*, 55–62. [CrossRef]
7. Futoma, J.; Simons, M.; Doshi-Velez, F.; Kamaleswaran, R. Generalization in Clinical Prediction Models: The Blessing and Curse of Measurement Indicator Variables. *Crit. Care Explor.* **2021**, *3*, e0453. [CrossRef]
8. Liu, Z.; Khojandi, A.; Mohammed, A.; Li, X.; Chinthala, L.K.; Davis, R.L.; Kamaleswaran, R. HeMA: A hierarchically enriched machine learning approach for managing false alarms in real time: A sepsis prediction case study. *Comput. Biol. Med.* **2021**, *131*, 104255. [CrossRef]
9. Farzaneh, N.; Ansari, S.; Lee, E.; Ward, K.R.; Sjoding, M.W. Collaborative strategies for deploying artificial intelligence to complement physician diagnoses of acute respiratory distress syndrome. *npj Digit. Med.* **2023**, *6*, 62. [CrossRef]
10. Reamaroon, N.; Sjoding, M.W.; Gryak, J.; Athey, B.D.; Najarian, K.; Derksen, H. Automated detection of acute respiratory distress syndrome from chest X-rays using Directionality Measure and deep learning features. *Comput. Biol. Med.* **2021**, *134*, 104463. [CrossRef]
11. Singhal, L.; Garg, Y.; Yang, P.; Tabaie, A.; Wong, A.I.; Mohammed, A.; Chinthala, L.; Kadaria, D.; Sodhi, A.; Holder, A.L.; et al. eARDS: A multi-center validation of an interpretable machine learning algorithm of early onset Acute Respiratory Distress Syndrome (ARDS) among critically ill adults with COVID-19. *PLoS ONE* **2021**, *16*, e0257056. [CrossRef] [PubMed]
12. Sjoding, M.W.; Taylor, D.; Motyka, J.; Lee, E.; Co, I.; Claar, D.; McSparron, J.I.; Ansari, S.; Kerlin, M.P.; Reilly, J.P.; et al. Deep learning to detect acute respiratory distress syndrome on chest radiographs: A retrospective study with external validation. *Lancet Digit. Health* **2021**, *3*, e340–e348. [CrossRef] [PubMed]
13. Sjoding, M.W.; Hofer, T.P.; Co, I.; Courey, A.; Cooke, C.R.; Iwashyna, T.J. Interobserver Reliability of the Berlin ARDS Definition and Strategies to Improve the Reliability of ARDS Diagnosis. *Chest* **2018**, *153*, 361–367. [CrossRef] [PubMed]
14. Bellani, G.; Pham, T.; Laffey, J.G. Missed or delayed diagnosis of ARDS: A common and serious problem. *Intensive Care Med.* **2020**, *46*, 1180–1183. [CrossRef]
15. Brady, A.P. Error and discrepancy in radiology: Inevitable or avoidable? *Insights Imaging* **2017**, *8*, 171–182. [CrossRef]
16. Busardò, F.P.; Frati, P.; Santurro, A.; Zaami, S.; Fineschi, V. Errors and malpractice lawsuits in radiology: What the radiologist needs to know. *Radiol. Med.* **2015**, *120*, 779–784. [CrossRef]
17. Song, X.; Weister, T.J.; Dong, Y.; Kashani, K.B.; Kashyap, R. Derivation and Validation of an Automated Search Strategy to Retrospectively Identify Acute Respiratory Distress Patients Per Berlin Definition. *Front. Med.* **2021**, *8*, 614380. [CrossRef]
18. Honavar, S. Electronic medical records—The good, the bad and the ugly. *Indian J. Ophthalmol.* **2020**, *68*, 417–418. [CrossRef]
19. Maley, J.H.; Thompson, B.T. Embracing the Heterogeneity of ARDS. *Chest* **2019**, *155*, 453–455. [CrossRef]
20. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 590–597. [CrossRef]

21. Johnson, A.E.W.; Pollard, T.J.; Berkowitz, S.J.; Greenbaum, N.R.; Lungren, M.P.; Deng, C.-y.; Mark, R.G.; Horng, S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **2019**, *6*, 317. [CrossRef] [PubMed]
22. Holste, G.; Wang, S.; Jiang, Z.; Shen, T.C.; Shih, G.; Summers, R.M.; Peng, Y.; Wang, Z. Long-Tailed Classification of Thorax Diseases on Chest X-ray: A New Benchmark Study. In *Data Augmentation, Labelling, and Imperfections, Proceedings of the Second MICCAI Workshop, DALI 2022, Held in Conjunction with MICCAI 2022, Singapore, 22 September 2022*; Springer: Cham, Switzerland, 2022; Volume 13567. [CrossRef]
23. Vardhan, A.; Makhnevich, A.; Omprakash, P.; Hirschorn, D.; Barish, M.; Cohen, S.L.; Zanos, T.P. A radiographic, deep transfer learning framework, adapted to estimate lung opacities from chest X-rays. *Bioelectron. Med.* **2023**, *9*, 1. [CrossRef] [PubMed]
24. Makhnevich, A.; Sinvani, L.; Cohen, S.L.; Feldhamer, K.H.; Zhang, M.; Lesser, M.L.; McGinn, T.G. The Clinical Utility of Chest Radiography for Identifying Pneumonia: Accounting for Diagnostic Uncertainty in Radiology Reports. *Am. J. Roentgenol.* **2019**, *213*, 1207–1212. [CrossRef] [PubMed]
25. Makhnevich, A.; Sinvani, L.; Feldhamer, K.H.; Zhang, M.; Richardson, S.; McGinn, T.G.; Cohen, S.L. Comparison of Chest Radiograph Impressions for Diagnosing Pneumonia: Accounting for Categories of Language Certainty. *J. Am. Coll. Radiol.* **2022**, *19*, 1130–1137. [CrossRef] [PubMed]
26. Karimi, D.; Dou, H.; Warfield, S.K.; Gholipour, A. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Med. Image Anal.* **2020**, *65*, 101759. [CrossRef]
27. Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-rays with Deep Learning. *arXiv* **2017**, arXiv:1711.05225.
28. Darapaneni, N.; Ranjan, A.; Bright, D.; Trivedi, D.; Kumar, K.; Kumar, V.; Paduri, A.R. Pneumonia Detection in Chest X-rays using Neural Networks. *arXiv* **2022**, arXiv:2204.03618.
29. Ukwuoma, C.C.; Qin, Z.; Heyat, M.B.B.; Akhtar, F.; Bamisile, O.; Muaad, A.Y.; Addo, D.; Al-antari, M.A. A hybrid explainable ensemble transformer encoder for pneumonia identification from chest X-ray images. *J. Adv. Res.* **2023**, *48*, 191–211. [CrossRef]
30. Zhang, J.; Xie, Y.; Pang, G.; Liao, Z.; Verjans, J.; Li, W.; Sun, Z.; He, J.; Li, Y.; Shen, C.; et al. Viral Pneumonia Screening on Chest X-ray Images Using Confidence-Aware Anomaly Detection. *IEEE Trans. Med. Imaging* **2021**, *40*, 879–890. [CrossRef]
31. Showkatian, E.; Salehi, M.; Ghaffari, H.; Reiazi, R.; Sadighi, N. Deep learning-based automatic detection of tuberculosis disease in chest X-ray images. *Pol. J. Radiol.* **2022**, *87*, 118–124. [CrossRef]
32. Xu, T.; Yuan, Z. Convolution Neural Network With Coordinate Attention for the Automatic Detection of Pulmonary Tuberculosis Images on Chest X-rays. *IEEE Access* **2022**, *10*, 86710–86717. [CrossRef]
33. Nabulsi, Z.; Sellergren, A.; Jamshy, S.; Lau, C.; Santos, E.; Kiraly, A.P.; Ye, W.; Yang, J.; Pilgrim, R.; Kazemzadeh, S.; et al. Deep learning for distinguishing normal versus abnormal chest radiographs and generalization to two unseen diseases tuberculosis and COVID-19. *Sci. Rep.* **2021**, *11*, 15523. [CrossRef]
34. Chowdhury, M.E.; Rahman, T.; Khandakar, A.; Mazhar, R.; Kadir, M.A.; Mahbub, Z.B.; Islam, K.R.; Khan, M.S.; Iqbal, A.; Emadi, N.A.; et al. Can AI Help in Screening Viral and COVID-19 Pneumonia? *IEEE Access* **2020**, *8*, 132665–132676. [CrossRef]
35. Oh, Y.; Park, S.; Ye, J.C. Deep Learning COVID-19 Features on CXR Using Limited Training Data Sets. *IEEE Trans. Med. Imaging* **2020**, *39*, 2688–2700. [CrossRef] [PubMed]
36. Yamac, M.; Ahishali, M.; Degerli, A.; Kiranyaz, S.; Chowdhury, M.E.; Gabbouj, M. Convolutional Sparse Support Estimator-Based COVID-19 Recognition from X-ray Images. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 1810–1820. [CrossRef]
37. Liu, F.; Zang, C.; Shi, J.; He, W.; Liang, Y.; Li, L. An Improved COVID-19 Lung X-ray Image Classification Algorithm Based on ConvNeXt Network. *Int. J. Image Graph.* **2023**, 2450036. [CrossRef]
38. Ozturk, T.; Talo, M.; Yildirim, E.A.; Baloglu, U.B.; Yildirim, O.; Acharya, U.R. Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Comput. Biol. Med.* **2020**, *121*, 103792. [CrossRef]
39. Rousan, L.A.; Elobeid, E.; Karrar, M.; Khader, Y. Chest X-ray findings and temporal lung changes in patients with COVID-19 pneumonia. *BMC Pulm. Med.* **2020**, *20*, 245. [CrossRef]
40. Gour, M.; Jain, S. Automated COVID-19 detection from X-ray and CT images with stacked ensemble convolutional neural network. *Biocybern. Biomed. Eng.* **2022**, *42*, 27–41. [CrossRef]
41. Khan, A.I.; Shah, J.L.; Bhat, M.M. CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest X-ray images. *Comput. Methods Programs Biomed.* **2020**, *196*, 105581. [CrossRef]
42. Zhong, A.; Li, X.; Wu, D.; Ren, H.; Kim, K.; Kim, Y.; Buch, V.; Neumark, N.; Bizzo, B.; Tak, W.Y.; et al. Deep metric learning-based image retrieval system for chest radiograph and its clinical applications in COVID-19. *Med. Image Anal.* **2021**, *70*, 101993. [CrossRef] [PubMed]
43. Saha, P.; Sadi, M.S.; Islam, M.M. EMCNet: Automated COVID-19 diagnosis from X-ray images using convolutional neural network and ensemble of machine learning classifiers. *Inform. Med. Unlocked* **2021**, *22*, 100505. [CrossRef] [PubMed]
44. Nahiduzzaman, M.; Goni, M.O.F.; Hassan, R.; Islam, M.R.; Syfullah, M.K.; Shahriar, S.M.; Anower, M.S.; Ahsan, M.; Haider, J.; Kowalski, M. Parallel CNN-ELM: A multiclass classification of chest X-ray images to identify seventeen lung diseases including COVID-19. *Expert Syst. Appl.* **2023**, *229*, 120528. [CrossRef] [PubMed]
45. Nasser, A.A.; Akhloufi, M.A. A Review of Recent Advances in Deep Learning Models for Chest Disease Detection Using Radiography. *Diagnostics* **2023**, *13*, 159. [CrossRef] [PubMed]
46. Ismael, A.M.; Şengür, A. Deep learning approaches for COVID-19 detection based on chest X-ray images. *Expert Syst. Appl.* **2021**, *164*, 114054. [CrossRef]

47. Singh, S.; Rawat, S.S.; Gupta, M.; Tripathi, B.K.; Alanzi, F.; Majumdar, A.; Khuwuthyakorn, P.; Thinnukool, O. Deep Attention Network for Pneumonia Detection Using Chest X-ray Images. *Comput. Mater. Contin.* **2023**, *74*, 1673–1691. [CrossRef]
48. Yuan, Z.; Yan, Y.; Sonka, M.; Yang, T. Large-scale Robust Deep AUC Maximization: A New Surrogate Loss and Empirical Studies on Medical Image Classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021.
49. Zhao, J.; Li, M.; Shi, W.; Miao, Y.; Jiang, Z.; Ji, B. A deep learning method for classification of chest X-ray images. *J. Phys. Conf. Ser.* **2021**, *1848*, 012030. [CrossRef]
50. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
51. Yao, L.; Poblenz, E.; Dagunts, D.; Covington, B.; Bernard, D.; Lyman, K. Learning to diagnose from scratch by exploiting dependencies among labels. *arXiv* **2017**, arXiv:1710.10501
52. Islam, M.Z.; Islam, M.M.; Asraf, A. A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images. *Inform. Med. Unlocked* **2020**, *20*, 100412. [CrossRef]
53. Guan, Q.; Huang, Y. Multi-label chest X-ray image classification via category-wise residual attention learning. *Pattern Recognit. Lett.* **2020**, *130*, 259–266. [CrossRef]
54. Heidari, M.; Mirniaharikandehei, S.; Khuzani, A.Z.; Danala, G.; Qiu, Y.; Zheng, B. Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms. *Int. J. Med. Inform.* **2020**, *144*, 104284. [CrossRef] [PubMed]
55. Huang, J.; Yang, S.; Wang, X. Enhancement Guidance Network for Classification of Pneumonia in Chest X-rays. In Proceedings of the 2022 4th International Conference on Robotics, Intelligent Control and Artificial Intelligence, Dongguan China, 16–18 December 2022; pp. 940–945. [CrossRef]
56. Motamed, S.; Rogalla, P.; Khalvati, F. Data augmentation using Generative Adversarial Networks (GANs) for GAN-based detection of Pneumonia and COVID-19 in chest X-ray images. *Inform. Med. Unlocked* **2021**, *27*, 100779. [CrossRef] [PubMed]
57. Gulakala, R.; Markert, B.; Stoffel, M. Rapid diagnosis of Covid-19 infections by a progressively growing GAN and CNN optimisation. *Comput. Methods Programs Biomed.* **2023**, *229*, 107262. [CrossRef] [PubMed]
58. Chambon, P.; Bluethgen, C.; Delbrouck, J.B.; der Sluijs, R.V.; Połacin, M.; Chaves, J.M.Z.; Abraham, T.M.; Purohit, S.; Langlotz, C.P.; Chaudhari, A. RoentGen: Vision-Language Foundation Model for Chest X-ray Generation. *arXiv* **2022**, arXiv:2211.12737.
59. Pai, K.C.; Chao, W.C.; Huang, Y.L.; Sheu, R.K.; Chen, L.C.; Wang, M.S.; Lin, S.H.; Yu, Y.Y.; Wu, C.L.; Chan, M.C. Artificial intelligence–aided diagnosis model for acute respiratory distress syndrome combining clinical data and chest radiographs. *Digit. Health* **2022**, *8*, 1–15. [CrossRef]
60. Reamaroon, N.; Sjoding, M.W.; Lin, K.; Iwashyna, T.J.; Najarian, K. Accounting for label uncertainty in machine learning for detection of acute respiratory distress syndrome. *IEEE J. Biomed. Health Inform.* **2019**, *23*, 407–415. [CrossRef] [PubMed]
61. Kim, Y.G.; Kim, K.; Wu, D.; Ren, H.; Tak, W.Y.; Park, S.Y.; Lee, Y.R.; Kang, M.K.; Park, J.G.; Kim, B.S.; et al. Deep Learning-Based Four-Region Lung Segmentation in Chest Radiography for COVID-19 Diagnosis. *Diagnostics* **2022**, *12*, 101. [CrossRef]
62. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
63. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
64. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
65. Müller, R.; Kornblith, S.; Google, G.H.; Toronto, B. When Does Label Smoothing Help? In Proceedings of the Advances in Neural Information Processing Systems 32 (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019.
66. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929
67. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
68. Glas, A.S.; Lijmer, J.G.; Prins, M.H.; Bonsel, G.J.; Bossuyt, P.M. The diagnostic odds ratio: A single indicator of test performance. *J. Clin. Epidemiol.* **2003**, *56*, 1129–1135. [CrossRef]
69. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626. [CrossRef]
70. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In *Computer Vision—ECCV 2014, Proceedings of the 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Springer: Cham, Switzerland, 2014.
71. Eyre, H.; Chapman, A.B.; Peterson, K.S.; Shi, J.; Alba, P.R.; Jones, M.M.; Box, T.L.; DuVall, S.L.; Patterson, O.V. Launching into clinical space with medspaCy: A new clinical text processing toolkit in Python. In Proceedings of the 2021 Annual Symposium, San Diego, CA, USA, 30 October–3 November 2021.
72. Guérin, C.; Reignier, J.; Richard, J.C.; Beuret, P.; Gacouin, A.; Boulain, T.; Mercier, E.; Badet, M.; Mercat, A.; Baudin, O.; et al. Prone Positioning in Severe Acute Respiratory Distress Syndrome. *N. Engl. J. Med.* **2013**, *368*, 2159–2168. [CrossRef] [PubMed]

73. The Acute Respiratory Distress Syndrome Network. Ventilation with Lower Tidal Volumes as Compared with Traditional Tidal Volumes for Acute Lung Injury and the Acute Respiratory Distress Syndrome. *N. Engl. J. Med.* **2000**, *342*, 1301–1308. [CrossRef] [PubMed]
74. Bellani, G.; Laffey, J.G.; Pham, T.; Madotto, F.; Fan, E.; Brochard, L.; Esteban, A.; Gattinoni, L.; Bumbasirevic, V.; Piquilloud, L.; et al. Noninvasive Ventilation of Patients with Acute Respiratory Distress Syndrome. Insights from the LUNG SAFE Study. *Am. J. Respir. Crit. Care Med.* **2017**, *195*, 67–77. [CrossRef]
75. Ganapathy, A.; Adhikari, N.K.; Spiegelman, J.; Scales, D.C. Routine chest X-rays in intensive care units: A systematic review and meta-analysis. *Crit. Care* **2012**, *16*, R68. [CrossRef] [PubMed]
76. Socha, M.; Prażuch, W.; Suwalska, A.; Foszner, P.; Tobiasz, J.; Jaroszewicz, J.; Gruszczynska, K.; Sliwinska, M.; Nowak, M.; Gizycka, B.; et al. Pathological changes or technical artefacts? The problem of the heterogenous databases in COVID-19 CXR image analysis. *Comput. Methods Programs Biomed.* **2023**, *240*, 107684. [CrossRef] [PubMed]
77. DeGrave, A.J.; Janizek, J.D.; Lee, S.I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* **2021**, *3*, 610–619. [CrossRef]