International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 01 | January - 2024 SJIF Rating: 8.176 ISSN: 2582-3930

# Analysis of Acute Respiratory Distress Syndrome using Machine Learning

**Harika Koheda, Dr. Vinoda Reddy**

Assistant Professor, Department of CSE, Malla Reddy Institute of Technology and Science (Autonomous), Hyderabad.

Associate Professor, Dept. of CSE(AIML), CMR Technical Campus, Hyderabad.

**Abstract:** Acute respiratory distress syndrome (ARDS) is a common but not well known critical disease condition that is linked to a high death rate. The fact that chest x-rays for ARDS are not always interpreted the same way is a big reason why it is not recognised more often. We wanted to teach a deep convolutional neural network (CNN) how to find signs of ARDS on chest x-rays. CNNs were first trained on 595,506 chest x-rays from two places to find common chest results like opacity and effusion. They were then trained on 8072 chest x-rays that had been marked for ARDS by several doctors using different transfer learning methods. The best CNN was tried on chest x-rays from both an internal and external group. Six doctors, including a chest radiologist and intensive care medicine specialists, looked over a subset of the images. Four hospitals in the US provided chest x-ray statistics. A CNN was able to find ARDS in 1560 chest x-rays from 455 patients with acute hypoxemic respiratory failure, with an area under the receiver operator characteristics curve (AUROC) of 0.92% (95% CI 0.89% to 0.94%). Its AUROC was 0.933 (95% CI 0.888–0.996) for the subset of 413 pictures looked at by at least six doctors, and its sensitivity was 83% (95% CI 74.0% to 91.1%). The AUROC was 0.93% (0.92–0.95%) when images marked as "equivocal" were taken out of the analysis. Chest x-rays can be used to train a CNN to perform as well as a doctor at finding ARDS. More study is needed to see how well these algorithms work for finding ARDS patients in real time so that evidence-based care is followed or to help with current ARDS research.

**Keywords:** Machine learning, support vector machine, label uncertainty, acute respiratory distress syndrome, sampling from longitudinal electronic health records (EHR).

## I.INTRODUCTION

Tachypnoea, severe hypoxemia, poor respiratory compliance, and damage to lung tissue seen on chest radiographs are the clinical hallmarks of acute respiratory distress syndrome (ARDS) [1]. Both ARDS and its milder variant, acute lung injury (ALI), are diagnosed via clinical characterization, even though diffuse alveolar destruction is the major pathological process [2]. Revisions to the clinical criteria for ALI/ARDS, often referred to as the "Berlin definition" [3], were made in 2012. Between 35% and 50% of all fatalities in the United States are attributable to acute respiratory distress syndrome (ARDS), which is also responsible for about 2 million critical care days and 75,000 deaths every year [4]. Once identified, ALI/ARDS often advances rapidly, and there are currently no specific and sensitive methods for early diagnosis. The survival percentage of patients with ARDS may be improved with early diagnosis and care, according to several scientific and clinical research [6]. Despite the equivalent importance, no models for forecasting ARDS episodes have been reported thus far. Consequently, a predictive model for ARDS occurrences is urgently required, since it has the potential to enhance the clinical diagnosis of ARDS.

An NIOSH definition from 2001 states that a biomarker is "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" [7]. The identification of ARDS may be aided by biomarkers, which reflect pathophysiological pathways. Therefore, it is possible to improve the diagnosis of ARDS by combining trustworthy biomarkers with current clinical classifications. Biomarkers have other potential uses beyond ARDS diagnosis, including risk stratification, outcome prediction, and surrogate endpoint monitoring [8]. The hunt for accurate ARDS biomarkers has been going on for the past twenty years, driven by the benefits of biomarkers [8] and the limitations of the American-European Consensus Criteria (AECC) criteria [9, 10]. Various biomarkers have been identified for the identification of acute respiratory distress syndrome (ARDS), including RAGE, Ang-2, SP-D, and inflammatory factors such as IL-6, IL-8, and TNF-α. (11, 12). So far, however, there are no clinical indicators that may be considered both sensitive and specific for ARDS [13].

Finding novel biomarkers that vary from the previously investigated biomarkers for ARDS and establishing a solid prediction model for ARDS events that incorporates these new biomarkers were the main aims of this secondary analysis of a prospective and independent cohort research.

## II. RELATED WORK

This research introduces the first prediction model for ARDS incidents, which includes eleven variables. Maximum and minimum heart rates and respiratory rates, minimum systolic blood pressures, temperature, white blood cell counts, and glucose, haematocrit, and salt levels were among the eleven variables that might be predicted. Additionally, there was a statistically significant association between ARDS occurrences and the lowest systolic blood pressure and highest and lowest respiratory rates on the first day of admission. The prediction model for ARDS occurrences was also updated to incorporate four additional biomarkers: lowered minimum haematocrit, glucose, and salt levels, and increased minimum WBC count. This was a first.

Lung inflammation may cause the potentially fatal acute respiratory distress syndrome. Despite over 30 completed or continuing clinical studies, no validated pharmacologic therapy exists for this illness, even if mechanical breathing strategies may alter death rates. Nevertheless, a number of research have shown that there are several predictors of death in ARDS patients, and numerous studies have revealed various prediction models for in-hospital mortality in ARDS patients. Twenty biomarkers for ARDS diagnosis and nineteen biomarkers for ARDS patient death prediction were published by Terpstra et al. [12]. Also, integrating numerous biomarkers may improve diagnosis accuracy, according to certain research. We developed a model to predict the occurrence of ARDS in intensive care unit patients in this research.

Our study's prediction model of ARDS episodes included eleven major predictors chosen from 42 variables. Our chosen predictors are biochemical markers of ARDS episodes, but prior research has shown that inflammatory factors and lung surface proteins are the most important determinants in ARDS patient death and diagnosis.

In addition, we added four fundamental vital signs in the model that predicts the occurrence of ARDS events. We discovered that critical patients with or without ARDS had greater minimum and maximum respiratory rates than healthy patients, and that ARDS patients had higher rates than non-ARDS patients. Also, in line with the clinical signs of ARDS, critical care patients with or without the disease had lower minimum systolic pressure and MAP than healthy patients, and ARDS patients had lower MAP than non-ARDS patients.

To top it all off, no other model has ever included these four novel biomarkers as potential ARDS event predictors.

## III. METHODS

Two publicly accessible datasets for chest radiographs, CheXpert13 and MIMIC-CXR, were merged to form the pretraining dataset.14 In the past, radiologists used a natural language processing algorithm on the reports they wrote for these pictures to annotate them for 14 typical clinical abnormalities detected on chest radiographs (excluding ARDS).

Patients who were admitted to the University of Michigan hospital between January 1, 2016, and June 30, 2017, and developed acute hypoxemic respiratory failure (a PaO2/FiO2 level of less than 300) while on one of the following respiratory support methods was included in the training dataset. These were invasive mechanical ventilation, non-invasive ventilation, or a heated high-flow nasal cannula. In the medical, surgery, cardiac, or neurological urgent care unit, patients got care. Patients who came to the University of Michigan from other hospitals were not included because they might have had ARDS before they came. It was decided that 80% of the training sample would be used for CNN training and 20% would be used for confirmation.

During the first seven days in the hospital, all chest x-rays were used for training. At least two doctors who are trained in critical care medicine and are interested in ARDS study looked at each chest x-ray on their own to see if it showed signs of ARDS. On a scale from 1 to 8, doctors rated each picture to see if it had bilateral opacities that were consistent with ARDS. They also looked at other clinical statistics from each patient's stay in the hospital. The rating went from 1 (high confidence, no ARDS) to 8 (high confidence, ARDS; see appendix p. 5). We used an eight-point scale to make the annotations as reliable as possible.15 It was hard for annotators to decide if a radiograph was consistent with ARDS because the eight-point scale didn't have a middle number. They had to do this while still expressing their confusion. There was a 0.56% intraclass association among doctors who looked at the same picture.

The internal testing dataset consisted of all straight patients who were brought to the University of Michigan hospital between July 1, 2017, and December 1, 2017, and who experienced acute hypoxemic respiratory failure, as explained above. The internal training set and the test set did not have any patients that were the

same. The criteria for including and excluding were the same as for the training dataset. The only difference was that only chest x-rays taken of patients who had acute hypoxemic respiratory failure were included to make the review more useful in the real world. Some doctors made notes on both the internal training and test sets of chest x-rays. A two-class latent class model (ARDS or not ARDS) was used to combine what doctors said and decide what to name the x-rays. A model with three groups (ARDS, unclear, and not ARDS) was also looked at. Nine doctors looked over a group of 413 chest x-rays in the internal test set. At least six doctors, including a chest x-ray specialist, looked over each picture.

The external testing dataset was made up of patients who were brought to the Hospital of the University of Pennsylvania (Philadelphia, PA, USA) between January 1, 2015, and December 31, 2017. These patients were part of a prospective sepsis cohort study. The dataset had lung x-rays from the first five days of being admitted to the intensive care unit. These x-rays had already been marked for ARDS as part of the prospective study, but not in the same way as the University of Michigan datasets. Individual pulmonologists from the University of Pennsylvania who were trained in clinical study on ARDS marked the pictures as having ARDS, being unclear, or not having ARDS.18 Doctors marked pictures as "equivocal" if they thought it would be hard to classify because of other problems on the image or bad skill.18 Doctors who looked at all the files were not told what the CNN finding was.

## IV. CNN Algorithm

The training process and the different transfer learning methods that were tested. The CNN was trained with code that was taken from a source that was made public (aligholami/CheXpert-Keras). First, CNNs were taught to find 14 common diagnostic chest x-ray results on the pretraining dataset. These included oedema, infiltrate, and pleural effusion. The network settings were then fine-tuned to find ARDS in the training sample from the University of Michigan. We compared networks that were taught using different transfer learning methods. This limited the number of network factors that could be tweaked in different parts of the model. CNNs that didn't go through the pretraining step were also taught. In all, seven CNNs were taught. In the appendix (p. 7), there are more detailed specifics. We used values from the validation part of the training dataset to do Platt scaling on the CNN output to make the calibration better.

The University of Michigan used a private test set to see how well all seven CNNs trained during the study could tell the difference between chest x-rays that showed ARDS. The CNN with the largest area under the receiver operator characteristics curve (AUROC) was chosen to be studied further. After that, this CNN was used for all research, even tests by outside sources. The area under the precision-recall curve (AUPRC) was also used as an alternative measure of discrimination. It shows the trade-off between sensitivity and positive prediction value. We found CNN's sensitivity and specificity after setting its adjusted probability level to 0.5 for finding chest x-rays that were consistent with ARDS. We created CIs and ran statistical tests using block bootstrapping to deal with repeated measures by resampling at the patient level.

A group of 413 chest x-rays from the internal test set were looked at by more doctors so that the best CNN could be compared to individual doctors. In this group, nine doctors marked chest x-rays, and each of them looked at at least 120 pictures. To find out how well each doctor did, their comment was compared to a reference standard made from the average of five other doctors in this group of nine who looked at the same picture. We plotted the true positive rates (sensitivity) and false positive rates (specificity) for each doctor against CNN's receiver operator characteristics curve for the same group of patients. The precision (positive prediction value) and recall (sensitivity) of each doctor were put against the model's precision-recall curve.

After putting chest x-rays into groups based on how many doctors marked them as showing ARDS, a boxplot of CNN's ARDS chance predictions for each picture was made. Gradient-weighted class activation mapping (Grad-CAM) was used to see where CNN focused on in each picture when CNN identified photos as having ARDS.22 After putting pictures into groups based on how many ARDS annotations they had, Grad-CAM visualisations were used to look at the images with the highest ARDS chance to learn more about how CNN decisions are made.

As a secondary study in the University of Michigan test set, CNN ability was compared in groups of patients based on age, gender, race, and body mass index (BMI). A calibration plot was made and the intersection and slope were found to check CNN's calibration (appendix p. 12). Because doctors who annotated the University of Michigan dataset also marked when patients fit all the criteria for ARDS, the time from the start of ARDS to CNN recognition was measured to see if there might be any delays if the network was used in real life. There could be a delay if the CNN didn't find ARDS on the first chest x-ray that doctors confirmed had ARDS, but on a later chest x-ray. An exploratory study was also carried out to see how well CNN worked when the three-class latent model was used to group chest x-rays.

Chest x-rays in the external test set had already been marked as ARDS, inconclusive, or not ARDS. To find the best CNN in this set of data, both chest x-rays that were marked as "equivocal" or "not ARDS" were looked at as "not ARDS." In a second study, performance measures were found after chest x-rays that were marked as "equivocal" were taken out. A boxplot of ARDS odds was made by putting chest x-rays into groups based on these types of annotation.

## V.PERFOMACNE EVALUAITON

The table shows the demographics of each sample. The training set from the University of Michigan had 8072 chest x-rays from 1778 patients. Based on a doctor's review, 2665 (33%) of these images were consistent with ARDS. The internal test set from the University of Michigan had 1560 chest x-rays from 455 patients, and 438 (28%) of them showed signs of ARDS. The external test set had 958 chest x-rays from 431 patients, and 445 (46%) of them were consistent with ARDS based on a doctor's review. Pneumonia and illness from a cause other than the lungs were the most common things that put people at risk for ARDS.

The CNN that did the best (AUROC 0.92, 95% CI 0.89% to 0.94%) had the last convolutional block and the layers that came after it fine-tuned to find ARDS, while the others stayed the same after pretraining (appendix p 2; version ii). Somewhat higher AUROC for this CNN than a network with all parameters fine-tuned (AUROC 0.91, 95% CI 0.89% to 0.94%; appendix p 2; version iii), but the difference wasn't really that big (p=0.56). However, there was a big drop in performance between training and evaluation for this later network (AUROC 0·97–0·89; p<0.001), which suggests that it was too well fitted (appendix p 10). The success of CNNs that did not go through chest x-ray pretraining was lower (appendix p. 10). So, the CNN shown in the annex (p. 2; version ii) was used for all future analysis.

The CNN was tested on a sample of 413 chest x-rays from the University of Michigan's internal test set. The images were also looked at by more doctors so that they could be compared to images taken by different doctors (figure 1). This stronger reference standard was used to measure the AUROC, which was 0.933 (95% CI: 0.880 to 0.996). CNN was 83·0% sensitive (95% CI 74·0–91·1) and 88·3% specific (95% CI 83·1–92·8) when a 50% chance of ARDS was used as a cutoff. The AUPRC score was 79 (with a 95% confidence interval of 63 to 88). The AUROC was 0.922 (95% CI 0.87-9.96) when the chest doctor was

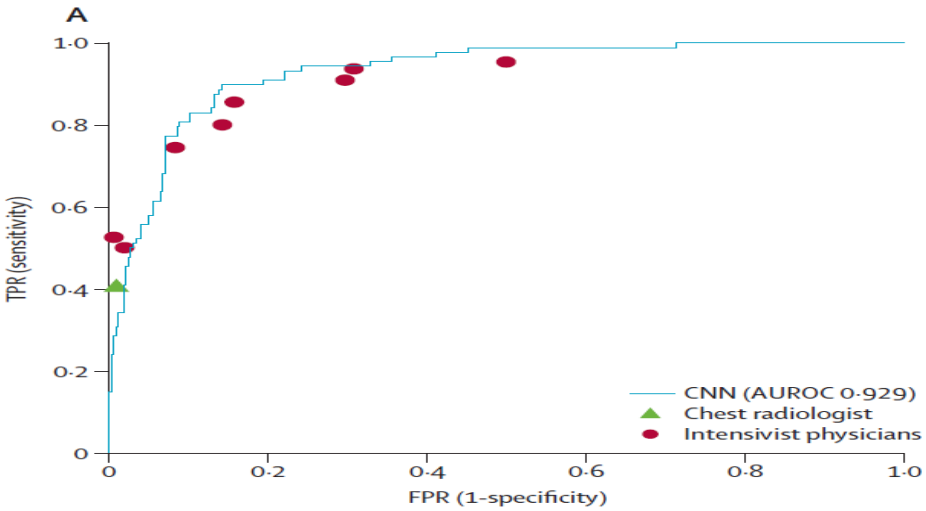
used as the only reference. The CNN performed about the same as each doctor. The doctors followed the CNN's receiver operator characteristic curve, either having better precision and lower sensitivity or the other way around. The CNN's estimate of the chance of ARDS matched the number of doctors who marked the x-ray as consistent with ARDS.

There were 155 chest x-rays with no more than six ARDS comments. The median calibrated CNN probability was 11%, and six (4%) of those 155 were given a chance above 50%. There were 27 chest x-rays with six ARDS comments. The median CNN probability was 91%, and only two (7%) of those 27 were given a chance below 50%. CNN gave middle-of-the-road odds to chest x-rays where doctors didn't agree on what they showed (for example, three out of six doctors marked the x-ray as showing ARDS).

| | | | |
|---|---|---|---|
| Patients | 1778 | 455 | 431 |
| Chest radiographs | 8072 | 1560 | 958 |
| Radiographs with ARDS* | 2665 (33%) | 438 (28%) | 445 (46%) |
| Age, years | 62 (51–71) | 63 (53–72) | 61 (52–69) |
| Sex | | | |
| Male | 1036 (58%) | 266 (58%) | 251 (58%) |
| Female | 742 (42%) | 189 (42%) | 180 (42%) |
| Race | | | |
| White | 1515 (85%) | 377 (83%) | 273 (63%) |
| Black | 164 (9%) | 49 (11%) | 129 (30%) |
| Other or unknown† | 99 (6%) | 29 (6%) | 29 (7%) |
| ARDS risk factor | | | |
| Pneumonia | 591 (33%) | 126 (28%) | 101 (23%) |
| Aspiration | 215 (12%) | 39 (9%) | NA |
| Non-pulmonary sepsis | 394 (22%) | 114 (25%) | 330 (77%) |
| Trauma | 110 (6%) | 28 (6%) | NA |
| APACHE score | 67 (52–85) | 68 (55–86) | 98 (72–129) |
| 30-day mortality | 420 (24%) | 119 (26%) | 188 (44%) |

Six of the six chest x-rays that were correctly labelled by doctors as showing ARDS had a high CNN likelihood. The CNN focused on parts of the lungs that showed opacities based on Grad-CAM images (figure 2). All six doctors marked the chest x-rays as showing ARDS, but the CNN gave them lower chances of finding something outside the lungs. When looking at chest x-rays that didn't have ARDS notes on them but had a higher CNN chance, the CNN focused on disease that was only on the right side. Finally, in a set of x-rays that caused doctors to argue but were given a higher CNN probability, the CNN focused on the right lung, which seemed to have more disease.

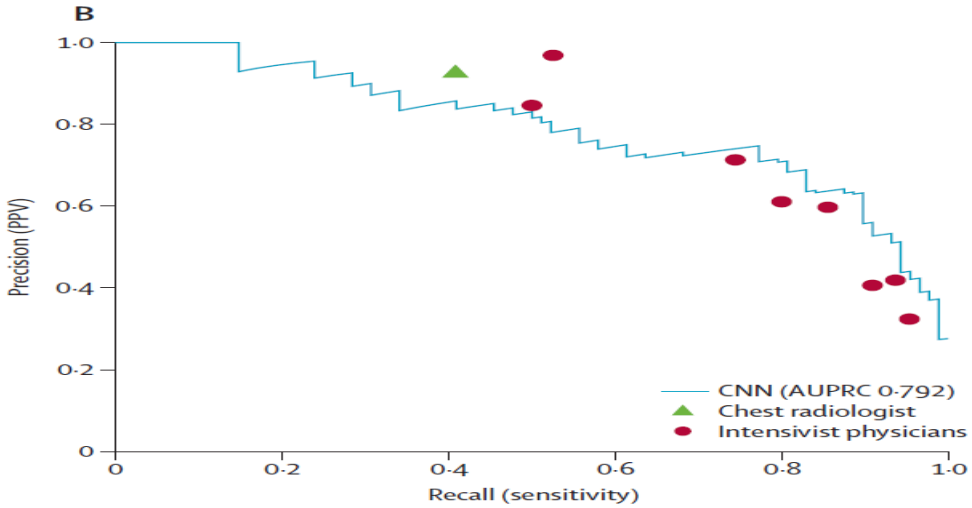
As a secondary study, CNN's performance was looked at across demographic groupings using the full University of Michigan internal test set (figure 3; appendix p 10). AUROC was the same for both men and women (0·91 vs. 0.93) (p=0.21). Plus, there wasn't a big difference between AUROC in White patients (0·92) and Black patients (0·90; p=0·63). Also, there wasn't a big difference between the age groups. The area under the curve (AUROC) was bigger in people whose BMI was between 30 and 35 kg/m² (0·96) than in people whose BMI was less than 25 kg/m² (0·89; p=0.004) or more than 35 kg/m² (0·96 vs 0·90; p=0.026). The CNN tuning is shown in the appendix (page 11). In a study to see if there would be a delay in finding patients with ARDS that CNN properly identified, the model found ARDS an average of 0 hours after doctors decided the patient met the criteria for ARDS, with a range of 4 hours before onset to 0 hours after onset. The average amount of time between a patient's first and last chest x-ray was 19 hours. This is an exploratory study that looked at CNN performance after putting chest x-rays into three groups. It found that CNN performed very well when chest x-rays in the "uncertain" class were taken out of the analysis.



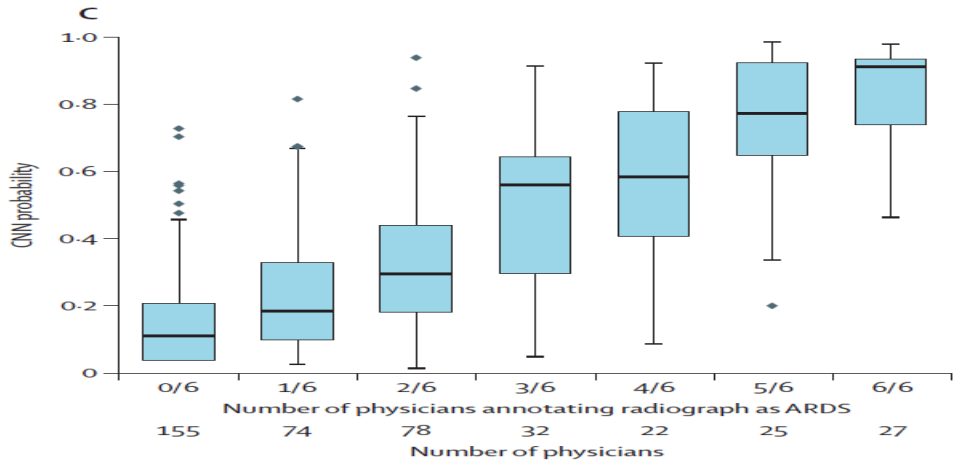

The CNN was then tested on an outside set of tests made by the University of Pennsylvania. AUPRC was 0·86 (95% CI 0·82–0·90; figure 4) and AUROC was 0·88 (95% CI 0·85–0·91). As an alternative study, chest x-rays marked as "equivocal" were not included. The AUROC was 0.93% (95% CI 0.92% to 0.95%) and the AUPRC was 0.95% (95% CI 0.92% to 0.96%). The CNN gave chest x-rays marked as "equivocal" a medium level of probability compared to chest x-rays marked as "ARDS" and those not marked as "ARDS." CNN tuning wasn't much worse than in the private test set.

When we used the CNN that was trained to find ARDS on chest x-rays from an outside centre, it only showed a small drop in accuracy. In other research that tried to train CNNs to find pneumonia, a network that was taught with data from two schools didn't work when it was tested at a third.24 Researchers thought that the network learned confusing features, like features of an image that let them tell which hospital system took the picture to see if it was consistent with pneumonia. One possible reason for our network's maintained success could be the limits that were put on it during transfer learning, which could have stopped it from overfitting. A two-center pretraining sample and expert physician comments, not radiology reports, were used to measure performance, which is another important difference.

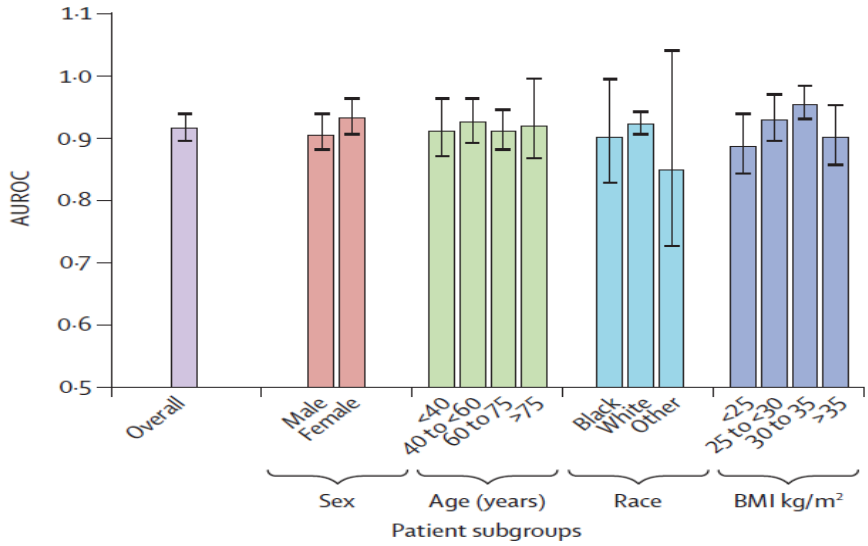
We made class activation maps (Grad-CAM) to see what the CNN was focused on when it decided that a scan showed ARDS. When CNN properly classified pictures, it seemed to focus on problems in the lungs. CNN sometimes got the labels of pictures wrong, making it look like they showed only one lung disease or finds outside the lungs. Some people think that CNNs should make choices based on the appearance of small local features in images, rather than how they are arranged spatially. Because of this, CNNs might find it hard to understand that local texture traits that look like lung damage must be in the lungs in order to represent ARDS. It's possible that networks that better take these links into account would work better. For example, segmenting the lungs together and then labelling abnormalities could help.




ARDS isn't always identified or isn't noticed right away in clinical practice. Also, patients don't always get the treatments that are suggested by guidelines, such as lung protective mechanical breathing and prone positions.Some people have suggested using automated warning systems to make it easier to spot ARDS26. This is because doctors are more likely to use evidence-based treatments when they know their patients have ARDS.




In the past, people who wanted to make automatic tools that could find ARDS would usually look at electronic health records and the text of radiology reports. The CNN created here, on the other hand, looks at digital chest x-rays directly.




However, setting up the network could require more money to be spent on health technology. For example, patients with a PaO2/FiO2 of less than 300 could need to be automatically identified from electronic health records. Radiographs stored in picture archiving and communications systems would also need to be analysed computationally, and doctors would need a way to get the results.

Our study has some flaws. For starters, ARDS is a syndrome with a set of similar clinical traits. Right now, the diagnosis of ARDS is based on a mix of clinical and radiological factors, and there isn't a gold standard test that is easy to get.28 In order to train a CNN to find signs of ARDS, comments from expert doctors were used, which is not always reliable.4 To solve this problem, we used a uniform scale and reference standard that were made up of reviews from several different doctors to make them more reliable. Second, the external test set was marked up using a different approach that was helpful for ARDS translational research but might not have been the best for evaluating algorithms. It also didn't have exact time stamps of when ARDS started, which meant that a possible discovery delay couldn't be tested. Still, the network seems strong because it did well in the external dataset, even though it had a different reference standard and different doctors who annotated it. Third, selection bias among patients in the training datasets might make it less useful in real life. The mix of patients studied was similar to that of other ARDS studies2, but there wasn't an equal number of men and women, and some types of patients (like those who had been traumatised) were underrepresented.

So, the network should be tested with more types of patients and in more hospital situations. Lastly, we tested the network using a 50% chance threshold to find ARDS. However, the actual threshold used to decide if a patient has ARDS is likely to depend on the situation. When providing lung protective ventilation, for example, a lower threshold to maximise sensitivity might be better. When recruiting patients for ARDS clinical studies, on the other hand, a higher threshold to maximise precision might be better.

Overall, these results show how powerful deep learning models are because they can be taught to correctly identify chest x-rays that show ARDS. More study is needed to see how these algorithms could help find ARDS patients in real time so that evidence-based care is followed or so that current ARDS research can be supported.

## V.CONCLUSION

Acute respiratory distress syndrome (ARDS) is a common critical illness condition that is linked to a high death rate but is not well understood in clinical practice. One important part of the Berlin ARDS diagnosis is that chest imaging must show bilateral airspace disease. However, doctors have very different ideas about what this means. Deep convolutional neural networks are a type of machine learning that has been taught to automatically find many useful results on pictures with the level of accuracy of a doctor. For example,

they can find diabetic retinopathy and skin cancer. In biological and clinical settings, we think our paper makes a big difference in how we solve standard classification problems. There is almost always some doubt when doctors try to figure out what's wrong with a patient. In a machine learning job, that diagnosis label could be used as the classification label or to guess what will happen. Usually, the diagnostic uncertainty that comes with the name is not taken into account when the model is being built. We show how the level of trust an expert doctor has in a diagnosis label can be used as important data in the training process for the model. Using the known error in diagnostics in the medical field is a method that can be used in many other medical situations as well. For instance, sepsis is a medical disease that needs to be caught early for the best patient care. But diagnostic doubt is common, which makes it harder to make strong methods for finding sepsis. Like ARDS, adding label uncertainty to the training of an algorithm for finding sepsis may make the algorithm work better. I think it would also be helpful to find better ways to add label uncertainty to machine learning models other than CNN, like neural networks and random forests. In clinical practice, doctors often make mistakes when making diagnoses, so these learning methods that account for label confusion could be very useful in other healthcare settings.

## REFERENCES

1. Bellani G, Laffey JG, Pham T, 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.
2. Weiss CH, Baker DW, Weiner S, et al. Low tidal volume ventilation use in acute respiratory distress syndrome. *Crit Care Med* 2016; 44: 1515–22.
3. Sjoding MW, Hofer TP, Co I, Courey A, Cooke CR, Iwashyna TJ. Interobserver reliability of the Berlin ARDS definition and strategies to improve the reliability of ARDS diagnosis. *Chest* 2018; 153: 361–67.
4. Peng JM, Qian CY, Yu XY, et al. Does training improve diagnostic accuracy and inter-rater agreement in applying the Berlin radiographic definition of acute respiratory distress syndrome? A multicenter prospective study. *Crit Care* 2017; 21: 12.
5. Goddard SL, Rubenfeld GD, Manoharan V, et al. The randomized educational acute respiratory distress syndrome diagnosis study: a trial to improve the radiographic diagnosis of acute respiratory distress syndrome. *Crit Care Med* 2018; 46: 743–48.

6. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016; 316: 2402–10.
7. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; 542: 115–18.
8. Nam JG, Park S, Hwang EJ, et al. Development and validation of deep learning-based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology* 2019; 290: 218–28.
9. Sun C, Shrivastava A, Singh S, Gupta A. Revisiting unreasonable effectiveness of data in deep learning era. 2017 IEEE International Conference on Computer Vision; Venice, Italy; Oct 22–29, 2017.
10. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? Proceedings of the 27th International Conference on Neural Information Processing Systems 2014; 2: 3320–28.
11. Huang G, Liu Z, Weinberger KQ. Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition; Honolulu, HI, USA; July 21–26, 2017: 2261–69.
12. Irvin J, Rajpurkar P, Ko M, et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. Proceedings of the AAAI Conference on Artificial Intelligence 2019; 33: 590–597.
13. Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR: a large publicly available database of labeled chest radiographs. *Sci Data* 2019; 6: 317.
14. Cicchetti DV, Shoinralter D, Tyrer PJ. The effect of number of rating scale categories on levels of interrater reliability: a Monte Carlo investigation. *Appl Psychol Meas* 1985; 9: 31–36.
15. Reilly JP, Wang F, Jones TK, et al. Plasma angiopoietin-2 as a potential causal marker in sepsis-associated ARDS development: evidence from Mendelian randomization and mediation analysis. *Intensive Care Med* 2018; 44: 1849–58.
16. Reilly JP, Meyer NJ, Shashaty MGS, et al. ABO blood type A is associated with increased risk of ARDS in whites following both major trauma and severe sepsis. *Chest* 2014; 145: 753–61.
17. Shah CV, Lanken PN, Localio AR, et al. An alternative method of acute lung injury classification for use in observational studies. *Chest* 2010; 138: 1054–61.
18. Gholami A, Chou B. CheXpert-Keras. 2019. https://github.com/ aligholami/CheXpert-Keras (accessed May 27, 2019).

19. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J Am Med Inform Assoc* 2020; 27: 621–33.
20. Field CA, Welsh AH. Bootstrapping clustered data. *J R Stat Soc Series B Stat Methodol* 2007; 69: 369–90.
21. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. 2017 IEEE International Conference on Computer Vision; Venice, Italy; Oct 22–29, 2017: 618–26.