ISBN 978-989-9121-31-7
2nd International Conference on Challenges in Engineering, Medical, Economics and Education: Research & Solutions (CEMEERS-23b)
Istanbul (Turkiye) Nov. 23-24, 2023

Benefits of Deep Learning Chest X-ray Features in the Mortality Prediction of COVID-19-associated ARDS Patients

Tiago Galvão[1], Ana Cysneiros[2,3], Pedro Jorge[1], Luís Bento[2,3], Nuno Domingues[1]

[1]Instituto Politécnico de Lisboa/ Instituto Superior de Engenharia de Lisboa
[2] Nova Medical School, Universidade de Lisboa
[3] Unidade de Urgência Médica, Hospital de São José, Centro Hospitalar Universitário Lisboa Central

***Abstract*:** *Acute respiratory distress syndrome associated with COVID-19 (ARDS-COV19) is a severe pulmonary syndrome leading to acute respiratory failure. ARDS is complex and heterogeneous, with patients frequently needing invasive mechanical ventilation (IMV) in intensive care units (ICUs). The identification of risk groups is crucial for precision medicine, although the lack of diagnostic methods can be limiting. Chest radiography is a qualitative and accessible imaging examination routinely used in ICU settings. The development of a multivariate and quantitative classifier based on radiomics is essential for predicting the mortality of patients under IMV and to evaluate the possible benefits of imaging features in the risk stratifications of this patients. For this purpose, 87 ARDS-COV19 patients from an ICU, with an average age of 64 ±11,6 years, of whom 62.07% were male, were included. The mortality rate was 48.28%. Chest X-rays from the 1st and 3rd days of IMV were collected, pre-processed, and concatenated. Deep learning features were then extracted using a pre-trained convolutional neural network (CheXnet). These features were combined with clinical variables (CV) to build four machine learning models based on logistic regression (LogReg) and multilayer perceptron (MLP) classifiers. Age, the PaO_2/FiO_2 ratio on the 3rd day of IMV, and a deep learning image feature (DLF) were selected by filter techniques to be used in two of the models. The remaining models were created only using the CV variables (Age, PaO2/FiO2). Using 10-fold cross validation the models which included the DLF_258 showed 89% (LogReg) and 82% (MLP) probability of having better accuracy than CV models with a 77% (MLP) and 86% (LogReg) chance of improved F1-Score. In the LogReg based models, AUC had a 74% chance of being improved with the added image features while on the MLP models only a 55% chance was noted. In cross-validation the combined LogReg models provided a mean AUC of 0.77, 95% CI [0.67, 0.85], an accuracy of 0.75, 95% CI [0.67, 0.85], and an F1 score of 0.73, 95% CI [0.62, 0.82], while the MLP models provided an AUC of 0.78, 95% CI [0.68, 0.86], an accuracy of 0.77, 95% CI [0.67, 0.85], and an F1 score of 0.76, 95% CI [0.67, 0.84]. Despite promising results, the sample size was limited, and external testing is lacking. Therefore, data collection and subsequent validation are essential.*

Keywords: *ARDS, ICU, COVID-19, X-ray, Mortality*

1. Introduction

The definition of ARDS has evolved over the years. The contemporary medical consensus resides on the Berlin definition, which relies on the assessment of medical images like chest radiographs (CXR) and arterial blood gas analysis. The evaluation of these gases allows for the assessment of the ratio between the partial pressure of oxygen and the inspired oxygen (PaO2/FiO2, or PF ratio), which, in turn, depends on the value of positive end-expiratory pressure (PEEP) for each respiratory cycle [1]. These values are essential in the current diagnosis and severity assessment of the described syndrome. ARDS can thus be defined as an acute pulmonary syndrome involving bilateral lung infiltrates (observable in imaging methods) and hypoxemia, not justifiable only by left ventricular heart dysfunction, fluid overload, or chronic lung disease [1], [2].

Despite its seemingly simple definition, ARDS is a syndrome of considerable complexity, as its expression and severity are co-associated with various pathologies and clinical disorders [3]–[5]. ARDS is a common, heterogeneous, and severe syndrome, regularly requiring invasive mechanical ventilation (IMV) in intensive care units (ICUs) [6]. Few therapeutic options have demonstrated benefits in patient prognosis beyond lung-protective ventilation, and this limitation may stem from the underlying heterogeneity of ARDS [6],[7].

The creation of new methods for identifying disease subphenotypes and classifying/assessing the severity of ARDS is thus essential and largely represents the current state of the art in achieving precision medicine. A subphenotype can be a subgroup of patients that present a higher risk of worse disease outcomes or share similar responses to medical procedures [4], [8], [5].

Recently, a new typology of ARDS has been considered, namely ARDS associated with COVID-19 (ARDS-COV19) [9]. The COVID-19 pandemic has caused increases in mortality and morbidity worldwide. Patients with COVID-19 are at a higher risk of requiring IMV and continue to be a current concern [6]. Recent research as shown benefits of using corticosteroids in certain patients with this type of infection, which has not been proven to work in common ARDS [10]. This demonstrates the possibility of subphenotypes associated with COVID-19 ARDS, proving suggesting that studying this syndrome in a homogenous population (COVID-19) may be beneficial with latent class analysis showing possible new subphenotypes [11], [12]. Various studies on protein biomarkers have shown a better understanding of ARDS and may offer a possible pathway to personalized approach and disease severity evaluation, however, acquisition of these biomarkers depends on an invasive procedure may present complications in patients undergoing IMV [13], [14].

Therefore, an alternative method to these invasive diagnostic procedures is necessary. The solution may lie in the untapped quantitative diagnostic potential of medical images, as in the case of chest radiography (CXR). Portable CXR is an economical, readily available diagnostic method used daily in ICUs for ARDS diagnosis and characterization, as well as assessing the severity COVID-19 [15]. CXR is strongly associated with ARDS, with its qualitative assessment forming the basis of its definition. Some studies have analysed chest imaging patterns and the severity of ARDS [15], [16]. The Radiographic Assessment of Lung Edema (RALE) score was developed as a semi-quantitative measure of lung edema in ARDS patients, with higher RALE scores being independently associated with lower PaO2/FiO2 ratios and worse overall survival. A change in RALE during the first days after the onset of ARDS is also independently associated with survival, however statistical significant association with mortality remains a challenge [17]. An analysis of ICU patients under invasive ventilation found that the RALE score has excellent diagnostic accuracy for ARDS, with an area under the receiver operating characteristic curve (AUC) of 0.91 [17], [18]. Despite the clear advantages, RALE remains a semi-quantitative method, depending on the subjective assessment of chest radiographs. This fact makes the method potentially time-consuming and subject to inter/intra-operator variabilities. Therefore, there is interest in an automatic method for quantitatively predicting mortality risk using chest radiographs to subsequently identify risk groups of the syndrome.

Machine learning in radiology is a growing research topic and has been successfully used in ARDS diagnosis with replicable quantitative computational methods. Radiomics methodologies allow the extraction of quantifiable features from medical images, either through classical image texture statistical methods or deep learning methods, for various classification, segmentation, or patient prognosis prediction tasks [19]–[21]. Studies focused on the categorization or prediction of ARDS severity and mortality use different angles, from biomarkers to lung physiology and chest imaging [12], [21]. However, the few studies combine clinical data and portable CXR, mostly focus on mortality prediction of COVID-19 ICU patients, not considering the ARDS definition in the selected population [22]– [24]. These studies show improved model performance metrics when using CXR image features, although they rarely provide the probability of this finding being true considering model generalization to different training groups, validation groups and sample sizes.

The present study aims to evaluate the possible benefits of using CXR deep learning image features (DLF) in COVID-19 ARDS mortality prediction by Bayesian statistics hybrid deep-learning and machine learning models.

2. Materials and Methods

2.1. Study Population

In collaboration with the pneumology medical team of the Thoracic Department at São José Hospital (North Lisbon University Hospital Center), a retrospective cohort study was conducted, considering all patients with ARDS-COVID-19 who underwent invasive mechanical ventilation (VMI) and were admitted to the Intensive Care Units (ICUs) of São José Hospital and Curry Cabral Hospital between April 2020 and January 2021.

The inclusion criteria for the study were confirmation of SARS-CoV-2 infection through nasopharyngeal PCR or bronchoalveolar lavage PCR, diagnosis of ARDS according to the Berlin definition, and the need for invasive mechanical ventilation. Exclusion criteria included age under 18, pregnancy, other contributing causes for ARDS such as trauma. Patients with poor-quality CXR and lack of ventilation data were also excluded. Data corresponds to the first 72 hours of ICU admission, during which SARS-CoV-2 was the sole isolated infectious agent. A total of 89 patients were selected, with two excluded due to R-RTX artifacts. These samples were used for training and validation of the classification models. In total, 87 patients were considered, with an mean age of 64±11.6 years (minimum: 26, maximum: 83). Male patients account for 62.07% (54) of the total. Of the 87 patients, 17,65% (15) had severe ARDS (PF ratio less than 100 mmHg) where 86,67% (13) were male. The majority of the patients had moderate ARDS (PF ratio less than or equal to 200 mmHg and more than 100 mmHg) corresponding to 64.71% (55) of the total patients, where 60% (33) were male. Mortality after 72 hours of ICU admission (including after hospital transfer) was 48.28% (42) in total, and this binary variable served as the target for model training (survivor/non-survivor).

2.2. Clinical Data

Data on age, gender, arterial blood gases, such as the PF ratio and $PaCO_2$ (CO_2), ventilation setting, such as PEEP and pressure support (PS), were collected. These variables were obtained during the first day of IVM (D1) and between 48 to 72 hours of the same (D3), resulting in a total of nine clinical variables/features.

2.3. Imaging Data

Portable AP (antero-posterior) incidence CXR were collected for each patient corresponding to D1 and D3. They were downloaded in PNG format with proper anonymization procedures, for a total of 174 images. CXR were pre-processed using the python library OpenCV. A gaussian blur filter with a (3x3) kernel was applied and contrast limited adaptive histogram equalization was also performed, using a Clip Limit of 2 and a Tile Grid Size of (8x8). These techniques provide noise reduction and image contrast optimization in the regions of interest, which have shown to improve accuracy in image classification tasks using neural networks [25]– [27]. Lung area was segmented with associated cropping to the region of interested. Experimentally, D1 and D3 CXR were vertically concatenated, resized and normalized for deep learning feature (DLF) extraction using convolutional neural networks (CNN) transfer learning.

A DenseNet-121 architecture loaded with the open source CheXNet weights was selected for this purpose. The CheXNet is a DenseNet-121 CNN first trained using the ChestX-ray14 dataset [28], [29]. This CNN provides accurate predictions of 14 thoracic pathologies from CXR including pneumonia. The original model consists of an end-to-end fine-tuned DenseNet-121 CNN (using the ImageNet weights). The same parameters were used to train the CNN with ChestX-ray14 dataset using an Adam optimization algorithm ($\beta 1 = 0.9$ and $\beta 2 = 0.999$). Mini- batches of size 16 were used and a learning rate of 0.001 decayed by a factor of 10 each time the validation loss plateaus after an epoch was also employed. After the mentioned pre-training and model-weights loading, the concatenated CXR served as inputs (previously normalized and resized to the required 224x224 size) [28]. Feature maps provided by the last convolutional layer, were extracted and flattened using average pooling,

resulting in 1024 CXR image features with no fine-tuning for our dataset employed. These features were combined and with the clinical data.

2.4. Model Building

A total of four binary classification models, using both logistic regression (LogReg) and multilayer perceptron (MLP) algorithms, were built on the *Orange: Data Mining software (Version: 3.36.0)*. The LogReg classifier used ridge regression (L2) and a C value of 2, while the MLP used 200 neurons in the hidden layer, with a ReLu activation function and an Adam optimizer. The non-survivor class was considered the positive class with a binary encoding of 1, while survivors were considered the negative class with a binary encoding of 0. Models were trained and validated using 10-fold cross validation (10-CV). In each fold, missing data was imputed by mean, data was normalized using z-score standardization and feature selection was performed in order to avoid overfitting. Filter methods were applied in feature selection using ANOVA feature ranking. The top ranking features were iteratively added until no significant improvement in ROC-AUC (area under the receiver operating characteristic curve) and accuracy occurred. In order to evaluate the usefulness of the CXR features, two model groups were created with both mentioned classifiers. Models A (LogReg_A and MLP_A), which considered both image and clinical features and Models B (LogReg_B and MLP_B) which considered only clinical data.

2.5. Statistical Analysis

Models were evaluated using binary classification performance metrics like AUC, accuracy (CA), sensitivity (Sens), specificity (Spec), precision (Prec) and F1 Score (F1). Cross validation model comparison was performed using the Bayesian interpretation of the t-test provided by the *Orange: Data Mining* software [30]. A negligible difference in performance of 5% was also considered. Computation of 95% confidence intervals (95% CI) for the proportional metrics was done through binomial exact methods (Clopper-Pearson) and using the *MedCalc Statistical Software version 22.014 (MedCalc Software bv, Ostend, Belgium*). DeLong's test for AUC comparison was used to obtain AUC 95% confidence with the same software [31]. Trained model classification interpretability and feature importance was analysed using Shapley Additive exPlanations values (SHAP Values) based on cooperative game theory [32]. Positive SHAP values positively impact the prediction where the magnitude represents the strength of the feature effect in the final classification.

3. Results

Optimal 10-CV performance metrics were found during feature selection by using 3 features in Models A and by using 2 features in Models B. During final model training (using the total of 87 patients) Models B best performing features consisted of patient age and D3 PF ratio (PF_D3), while Models A employed the addition of one image DLF to the mentioned clinical features.

Table 1 and Table 2 presents the 10-CV performance metrics with 95% CI for the LogReg and MLP models respectively, where P_AB = Probability of model A being better than B, P_BA = Probability of model B being better than model A. Considering a negligible 5% performance difference negligible: P5%_AB = Probability of model A being better than B, P5%_BA = Probability of model B being better than A, Pneg5% = Probability of the models being identical. AUC_p = ROC AUC for the positive class and AUC_{mean}= mean ROC AUC between the classes. Higher probabilities of the Bayesian t-test are in bold and 95%CI are presented in brackets.

Cross validation results for the LogReg_A model provided an AUC of 0.77, 95% CI [0.67, 0.85], an accuracy of 0.75, 95% CI [0.67, 0.85], and an F1 score of 0.73, 95% CI [0.62, 0.82], while the MLP_A model provided an AUC of 0.78, 95% CI [0.68, 0.86], an accuracy of 0.77, 95% CI [0.67, 0.85], and an F1 score of 0.76, 95% CI [0.67, 0.84]. Introducing imaging features during model training provided a 89% (LogReg) and 82% (MLP) chance of improving model accuracy and a 77% (MLP) and 86% (LogReg) chance of improving F1-Score.

Regarding the LogReg models, there was a 63% chance of improvement considering a 5% difference in accuracy negligeable and a 71% chance of improvement considering a 5% difference in F1-score negligeable. Thus, the model performance boost provided by the addition of imaging features may be superior to 5% considering this metrics. Figure 1 shows the obtained SHAP values in the training data and feature importance.

TABLE I: 10-CV LogReg performance metrics and their respective Bayesian t-test results.

Model	AUC_{mean}	AUC_p	CA	Sens	Prec	Spec	F1
LogReg_A	0.77 [0.67 ,0.85]	0.78 [0.67 ,0.85]	0.74 [0.64 , 0.83]	0.71 [0.55 , 0.84]	0.75 [0.58 , 0.87]	0.78 [0.62 , 0.88]	0.73 [0.62 ,0.82]
LogReg_B	0,70 [0.59 ,0.79]	0.75 [0.59 ,0.79]	0.67 [0.56 ,0.77]	0.64 [0.48 ,0.78]	0.68 [0.51 ,0.81]	0.71 [0.56 ,0.83]	0,66 [0.55 ,0.76]
Bayesian t-test	P_AB	**0.74**	**0.89**	**0.77**	**0.88**	**0.76**	**0.86**
	P_BA	0.26	0.11	0.23	0.12	0.24	0.14
	$P_{5\%}$_AB	0.36	**0.63**	**0.57**	**0.73**	**0.55**	**0.71**
	$P_{5\%}$_BA	0.06	0.03	0.10	0.204	0.11	0.06
	$P_{neg5\%}$	**0.57**	0.45	0.33	0.230	0.35	0.23

TABLE II: 10-CV MLP performance metrics and their respective Bayesian t-test results.

Model	AUC_{mean}	AUC_p	CA	Sens	Prec	Spec	F1
MLP_A	0.78 [0.68 ,0.86]	0.80 [0.68 ,0.86]	0.77 [0.67 ,0.85]	0.74 [0.58 ,0.86]	0,78 [0.61 ,0.89]	0.74 [0.58 ,0.86]	0.76 [0.67 , 0.84]
MLP_B	0,74 [0.63 ,0.83]	0.79 [0.63 ,0.83]	0.70 [0.59 ,0.79]	0,76 [0.61 ,0.88]	0.67 [0.52 ,0.80]	0.71 [0.56 ,0.83]	0.71 [0.66 ,0.80]
Bayesian t-test	P_AB	**0.55**	**0.82**	0.43	**0.87**	**0.88**	**0.77**
	P_BA	0.45	0.18	**0.57**	0.13	0.12	0.23
	$P_{5\%}$_AB	0.32	**0.61**	0.22	**0.73**	**0.79**	**0.52**
	$P_{5\%}$_BA	0.23	0.07	0.11	0.05	0.06	0.09
	$P_{neg5\%}$	**0.46**	0.32	**0.56**	0.22	0.15	0.39

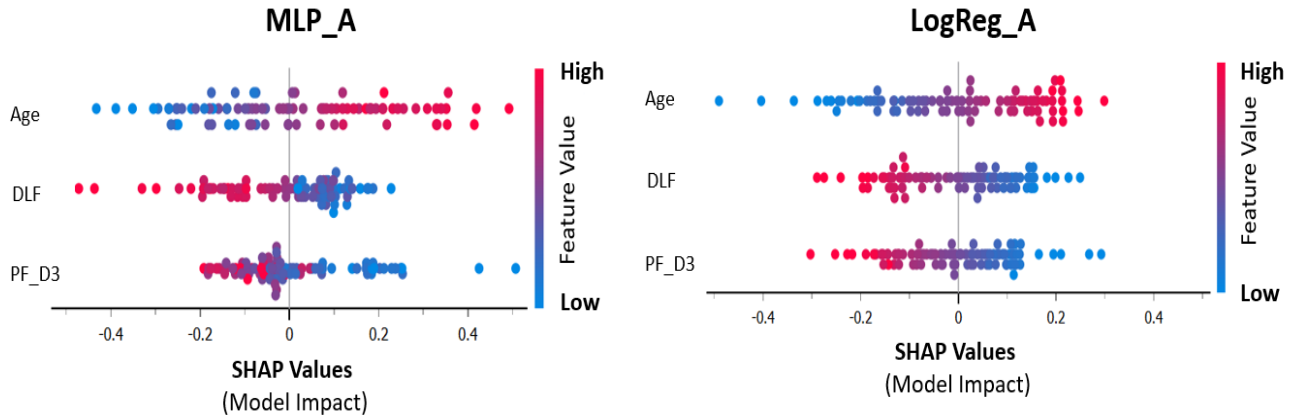

Figure 1 – SHAP Analysis of the training data (87 samples) predictions in models A (MLP and LogReg)

4. Discussion

Using transfer learning it was possible to extract 1024 CXR features for mortality prediction in COVID-19 patients with ARDS. This was done through image preprocessing (segmentation and contrast/noise optimization) and automatic extraction of deep learning features using a pre-trained CNN for CXR classification (CheXnet). The performance metrics results seem to suggest a performance boost when adding a CXR DLF to clinical data in model building which as shown to have a high probability of being superior to 5% in terms of accuracy, precision, sensitivity, specificity and F1 Score metrics (in the LogReg_A model).

Similar findings have been reported by D. Gourdeau et al. in 2022, when employed hybrid machine learning methodologies using a support vector machine. CRX DLF were extracted to predict mortality in COVID-19 patients undergoing IMV in ICU settings. A pre-trained CheXNet was also used for transfer learning and feature extraction purposes. The authors' most effective model involved a combination of a risk score associated with clinical variables and two CXR DLF. This model demonstrated an AUC of 0.74 95% CI[0.73, 0.75], a classification accuracy of 0.76, specificity of 0.83, and recall of 0.49 (the authors did not provide a 95% CI for recall), while the logistic regression model only using clinical variables presented an AUC of 0.620 [22].

J. Cheng et al. in 2022 also developed a mortality prediction method for COVID-19 patients admitted to the ICU. The authors employed a sophisticated deep learning architecture trained from scratch for the extraction of longitudinal features from segmented CXR in several days of IMV. Among the models created, those exhibiting greater similarity to the current study solely utilized radiographs during ICU admission and combined longitudinal ICU and pre-ICU CXR with clinical variables. In the first model, in an external test group, the authors reported an AUC of 0.70 95% CI [0.61, 0.78], a accuracy of 0.66 95% CI [0.58, 0.73], an F1 Score of 0.64 95% CI [0.55, 0.73], and a specificity 0.64 95% CI [0.55, 0.75]. The combined model, presented an AUC of 0.73 95% CI [0.65, 0.80], an accuracy of 0.73 95% CI [0.67, 0.81], an F1 score of 0.71 95% CI [0.62, 0.79], and a specificity of 0.75 95% CI [0.65, 0.83] [23]. Despite similarities in performance between both the mentioned reports and the current study, direct comparison should be approached with caution, given that the authors had a larger number of samples and an external test group, allowing for statistically robust results with narrower 95% confidence intervals in performance metrics. Although there is a likelihood that patients undergoing IMV in ICU may have ARDS-COV19, this condition was not explicitly mentioned by the authors, and there is no available data to support it according to the Berlin definition.

Even though an overall performance boost was verified in the LogReg_A model, performance varied between the used classification algorithms (MLP and LogReg). There was a low probability of performance increase in terms of AUC and sensitivity when considering more complex models like the used MLP (Table 2). Through the analysis of SHAP values of both models, it was found that the deep learning image feature was considered more important than the PF ratio in predicting patient mortality on the training dataset, indicating the potential importance of radiomics in ARDS. A possible sign of overfitting can also be found in the MLP model despite having better performance metrics than LogReg models. In SHAP analysis of the MLP model the age feature had considerable more impact in classification decision when compared to the other variables. which might explain the lack of improvement in AUC and sensitivity in this model with the addition of the DLF feature. On the other hand the LogReg model provides more balanced SHAP values between the features. This may be due to higher model complexity of the MLP and low sample training data.

The low sample dataset and lack of an external testing set was the main limitation of this study, introducing a higher risk of bias and variance and limiting the amount of features used to avoid potential overfitting. The fact that there was no fine-tuning done to the used CNN due to the limited amount of data also hinders the identification of what the deep learning feature represents through common techniques like Gradient-weighted Class Activation Mapping, producing a black-box algorithm. The input of two concatenated images into the CNN is also significantly different from the original images the CheXNet was originally trained on, with the possibility of producing unoptimized extracted DLF’s for the current task.

5. Conclusion

Due to the large volume of images produced daily in ICU care and their economic efficiency, chest radiography is an attractive proposal for investigating high risk patients with ARDS-COV19. The limitations in this study should serve as impetus for future work since the results provided by Bayesian statistics seem to suggest that the introduction of radiological information has a high probability of promoting the identification of risk groups in ARDS-COV19. The constructed models were able to achieve promising performance metrics in cross-validation using information available in any ICU, regardless of economic and professional resources,

although external test validation is still lacking. Future work could reside in fine-tuning an end-to-end CNN or exploring the use of hand-crafted radiomics features which may allow more interpretable results with less data. The combination of quantitative radiomics methods with daily clinical and biochemical assessment in ICUs may thus be the future of personalized medicine in the context of ARDS-COV19.

6. Acknowledgements

This research was funded by grant DSAIPA/DS/0117/2020, from FCT, Portugal.

7. References

[1] ARDS Definition Task Force, Ranieri VM, Rubenfeld GD, et al (2012) Acute respiratory distress syndrome: the Berlin Definition. JAMA 307:2526–33. https://doi.org/10.1001/jama.2012.5669

[2] Matthay MA, Zemans RL, Zimmerman GA, et al (2019) Acute respiratory distress syndrome. Nat Rev Dis Prim 5:18. https://doi.org/10.1038/s41572-019-0069-0

[3] Ware LB, Matthay MA (2000) The acute respiratory distress syndrome. N Engl J Med 342:1334–49. https://doi.org/10.1056/NEJM200005043421806

[4] Wildi K, Livingstone S, Palmieri C, et al (2021) Correction to: The discovery of biological subphenotypes in ARDS: a novel approach to targeted medicine? J Intensive Care 9:22. https://doi.org/10.1186/s40560-021-00534-y

[5] Maddali M V, Churpek M, Pham T, et al (2022) Validation and utility of ARDS subphenotypes identified by machine-learning models using clinical data: an observational, multicohort, retrospective analysis. Lancet Respir Med 10:367–377. https://doi.org/10.1016/S2213-2600(21)00461-6

[6] Krynytska I, Marushchak M, Birchenko I, et al (2021) COVID-19-associated acute respiratory distress syndrome versus classical acute respiratory distress syndrome (a narrative review). Iran J Microbiol 13:737–747. https://doi.org/10.18502/ijm.v13i6.8072

[7] Acute Respiratory Distress Syndrome Network, Brower RG, Matthay MA, et al (2000) 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 342:1301–8. https://doi.org/10.1056/NEJM200005043421801

[8] Shankar-Hari M, Rubenfeld GD (2019) Population enrichment for critical care trials: phenotypes and differential outcomes. Curr Opin Crit Care 25:489–497. https://doi.org/10.1097/MCC.0000000000000641

[9] Meyer NJ, Gattinoni L, Calfee CS (2021) Acute respiratory distress syndrome. Lancet 398:622–637. https://doi.org/10.1016/S0140-6736(21)00439-6

[10] RECOVERY Collaborative Group, Horby P, Lim WS, et al (2021) Dexamethasone in Hospitalized Patients with Covid-19. N Engl J Med 384:693–704. https://doi.org/10.1056/NEJMoa2021436

[11] Demšar J, Zupan B (2021) Hands-on training about overfitting. PLOS Comput Biol 17:e1008671. https://doi.org/10.1371/journal.pcbi.1008671

[12] Sinha P, Furfaro D, Cummings MJ, et al (2021) Latent Class Analysis Reveals COVID-19-related Acute Respiratory Distress Syndrome Subgroups with Differential Responses to Corticosteroids. Am J Respir Crit Care Med 204:1274–1285. https://doi.org/10.1164/rccm.202105-1302OC

[13] Villar J, Ferrando C, Martínez D, et al (2020) Dexamethasone treatment for the acute respiratory distress syndrome: a multicentre, randomised controlled trial. Lancet Respir Med 8:267–276. https://doi.org/10.1016/S2213-2600(19)30417-5

[14] Matthay MA, Arabi YM, Siegel ER, et al (2020) Phenotypes and personalized medicine in the acute respiratory distress syndrome. Intensive Care Med 46:2136–2152. https://doi.org/10.1007/s00134-020-06296-9

[15] Hui TCH, Khoo HW, Young BE, et al (2020) Clinical utility of chest radiography for severe COVID-19. 10:1540–1550. https://doi.org/10.21037/qims-20-642

[16] Zompatori M, Ciccarese F, Fasano L (2014) Overview of current lung imaging in acute respiratory distress syndrome. Eur Respir Rev 23:519–530. https://doi.org/10.1183/09059180.00001314

[17] Warren MA, Zhao Z, Koyama T, et al (2018) Severity scoring of lung e dema on the chest radiograph is associated with clinical outcomes in ARDS. Thorax 73:840–846. https://doi.org/10.1136/thoraxjnl-2017-211280

[18] Jabaudon M, Audard J, Pereira B, et al (2020) Early Changes Over Time in the Radiographic Assessment of Lung Edema Score Are Associated With Survival in ARDS. Chest 158:2394–2403. https://doi.org/10.1016/j.chest.2020.06.070

[19] Le S, Pellegrini E, Green-Saxena A, et al (2020) Supervised machine learning for the early prediction of acute respiratory distress syndrome (ARDS). J Crit Care 60:96–102. https://doi.org/10.1016/j.jcrc.2020.07.019

[20] Reamaroon N, Sjoding MW, Gryak J, et al (2021) Automated detection of acute respiratory distress syndrome from chest X-Rays using Directionality Measure and deep learning features. Comput Biol Med 134:104463. https://doi.org/10.1016/j.compbiomed.2021.104463

[21] Sjoding MW, Taylor D, Motyka J, et al (2021) Deep learning to detect acute respiratory distress syndrome on chest radiographs: a retrospective study with external validation. Lancet Digit Heal 3:e340–e348. https://doi.org/10.1016/S2589-7500(21)00056-X

[22] Gourdeau D, Potvin O, Biem JH, et al (2022) Deep learning of chest X-rays can predict mechanical ventilation outcome in ICU-admitted COVID-19 patients. Sci Rep 12:6193. https://doi.org/10.1038/s41598-022-10136-9

[23] Cheng J, Sollee J, Hsieh C, et al (2022) Correction to: COVID-19 mortality prediction in the intensive care unit with deep learning based on longitudinal chest X-rays and clinical data. Eur Radiol 32:5034–5034. https://doi.org/10.1007/s00330-022-08680-z

[24] Jiao Z, Choi JW, Halsey K, et al (2021) Prognostication of patients with COVID-19 using artificial intelligence based on chest x-rays and clinical data: a retrospective study. Lancet Digit Heal 3:e286–e294. https://doi.org/10.1016/S2589-7500(21)00039-X

[25] Giełczyk A, Marciniak A, Tarczewska M, Lutowski Z (2022) Pre-processing methods in chest X-ray image classification. PLoS One 17:e0265949. https://doi.org/10.1371/journal.pone.0265949

[26] Demircioğlu A (2022) Predictive performance of radiomic models based on features extracted from pretrained deep networks. Insights Imaging 13:187. https://doi.org/10.1186/s13244-022-01328-y

[27] Heidari M, Mirniaharikandehei S, Khuzani AZ, et al (2020) Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms. Int J Med Inform 144:104284. https://doi.org/10.1016/j.ijmedinf.2020.104284

[28] Rajpurkar P, Irvin J, Zhu K, et al (2017) CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. 3–9

[29] Wang X, Peng Y, Lu L, et al (2017) ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proc - 30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017 2017-Janua:3462–3471. https://doi.org/10.1109/CVPR.2017.369

[30] Corani G, Benavoli A (2015) A Bayesian approach for comparing cross-validated algorithms on multiple data sets. Mach Learn 100:285–304. https://doi.org/10.1007/s10994-015-5486-z

[31] DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44:837–845

[32] Lundberg S, Lee S-I (2017) A Unified Approach to Interpreting Model Predictions. Nips 16:426–430