Medical Image Analysis 74 (2021) 102216

Contents lists available at ScienceDirect

Medical Image Analysis

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

AIforCOVID: Predicting the clinical outcomes in patients with COVID-19 applying AI to chest-X-rays. An Italian multicentre study

Paolo Soda[a,*], Natascha Claudia D'Amico[a,b], Jacopo Tessadori[c], Giovanni Valbusa[d], Valerio Guarrasi[a,e], Chandra Bortolotto[f], Muhammad Usman Akbar[c,g], Rosa Sicilia[a], Ermanno Cordelli[a], Deborah Fazzini[b], Michaela Cellina[h], Giancarlo Oliva[h], Giovanni Callea[f], Silvia Panella[i], Maurizio Cariati[j], Diletta Cozzi[k], Vittorio Miele[k], Elvira Stellato[r], Gianpaolo Carrafiello[l,m], Giulia Castorani[n], Annalisa Simeone[o], Lorenzo Preda[f,p], Giulio Iannello[a], Alessio Del Bue[c], Fabio Tedoldi[d], Marco Alí[b,d], Diego Sona[c,q], Sergio Papa[b]

[a] Unit of Computer Systems and Bioinformatics, Department of Engineering, University Campus Bio-Medico of Rome, Via Alvaro del Portillo 21, Rome 00128, Italy
[b] Department of Diagnostic Imaging and Stereotactic Radiosurgey, Centro Diagnostico Italiano S.p.A., Via S. Saint Bon 20, Milan 20147, Italy
[c] Pattern Analysis and Computer Vision, Istituto Italiano di Tecnologia, Via Morego 30, Genoa 16163, Italy
[d] Bracco Imaging S.p.A., Via Caduti di Marcinelle 13, Milan 20134, Italy
[e] Department of Computer, Control, and Management Engineering, Sapienza University of Rome, Via Ariosto, 25, Rome 00185, Italy
[f] Radiology Institute, Fondazione IRCCS Policlinico San Matteo, Viale Golgi 19, Pavia 27100, Italy
[g] Department of Naval, Electrical, Electronic and Telecommunications Engineering University of Genova, Via All'Opera Pia 11 A, Genoa 16145, Italy
[h] Radiology Department, ASST Fatebenefratelli Sacco, Piazza Principessa Clotilde 3, Milan 20121, Italy
[i] Diagnostic and interventional radiology unit, ASST Santi Paolo e Carlo - San Paolo Hospital, Via Antonio di Rudiní 8, Milan 20142, Italy
[j] Department of Advanced Diagnostic Technologies - Therapeutic, Diagnostic and Radiology Units, ASST Santi Paolo e Carlo - San Paolo Hospital, Via Antonio di Rudiní 8, Milan 20142, Italy
[k] Department of Emergency Radiology, Careggi University Hospital, Largo Piero Palagi 1, Florence 50139, Italy
[l] Operative Unit of Radiology, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico of Milan, Via della Commenda, 10, Milan 20122, Italy
[m] Department of Health Sciences, Univeristy of Milan, Via Festa del Perdono, 7, Milan 20122, Italy
[n] Diagnostic Imaging, Postgraduate Medical School, University of Foggia, Via Antonio Gramsci 89, Foggia 71122, Italy
[o] Department of Diagnostic Imaging, IRCCS Ospedale Casa Sollievo della Sofferenza, Viale Cappuccini 1, San Giovanni Rotondo 71013, Italy
[p] Radiology Unit, Department of Clinical, Surgical, Diagnostic, and Pediatric Sciences, University of Pavia, Corso Str. Nuova, 65, Pavia 27100 Italy
[q] Fondazione Bruno Kessler, Via Sommarive, 18, Trento 38123, Italy
[r] Postgraduation School in Radiodiagnostics, Universitá degli Studi di Milano, Via Festa del Perdono, 7, Milan 20122, Italy

ARTICLE INFO

Article history:
Received 4 December 2020
Revised 3 August 2021
Accepted 18 August 2021
Available online 28 August 2021

Keywords:
COVID-19
Artificial intelligence
Deep learning
Prognosis

ABSTRACT

Recent epidemiological data report that worldwide more than 53 million people have been infected by SARS-CoV-2, resulting in 1.3 million deaths. The disease has been spreading very rapidly and few months after the identification of the first infected, shortage of hospital resources quickly became a problem. In this work we investigate whether artificial intelligence working with chest X-ray (CXR) scans and clinical data can be used as a possible tool for the early identification of patients at risk of severe outcome, like intensive care or death. Indeed, further to induce lower radiation dose than computed tomography (CT), CXR is a simpler and faster radiological technique, being also more widespread. In this respect, we present three approaches that use features extracted from CXR images, either handcrafted or automatically learnt by convolutional neuronal networks, which are then integrated with the clinical data. As a further contribution, this work introduces a repository that collects data from 820 patients enrolled in six Italian hospitals in spring 2020 during the first COVID-19 emergency. The dataset includes CXR images, several clinical attributes and clinical outcomes. Exhaustive evaluation shows promising performance both in 10-fold and leave-one-centre-out cross-validation, suggesting that clinical data and images have the potential to provide useful information for the management of patients and hospital resources.

* Corresponding author.
E-mail address: p.soda@unicampus.it (P. Soda).

https://doi.org/10.1016/j.media.2021.102216

1. Introduction

According to data reported by the European Centre for Disease Prevention and Control[1] as of 13 November 2020 almost 53 million patients worldwide have been infected with the new coronavirus SARS-CoV-2, causing 1.3 million deaths. Since the identification of patient zero in China, the situation dramatically worsened worldwide, saturating the healthcare system resources. With a shortage of beds available in intensive and sub-intensive care, the need for a quick and effective triage system became an urgency.

Chest imaging examinations, as chest X-ray (CXR) (Schiaffino et al., 2020) and computed tomography (CT) (Ai et al., 2020) play a pivotal role in different settings. Indeed, imaging is used during triage in case of unavailability, delay of or the first negative result of reverse transcriptase-polymerase chain reaction (RT-PCR) (Lu et al., 2020). Moreover, imaging is used to stratify disease severity. Generally, the findings on chest imaging in COVID-19 are not specific and overlap with other infections. CT should not be used to screen for or as a first-line test to diagnose COVID-19 and should be used sparingly and reserved for hospitalized, symptomatic patients with specific clinical indications for CT (American College of Radiology, 2020). CXR most frequent lesions in COVID-19 patients are reticular alteration (up to 5 days from the symptoms onset), and ground-glass opacity (after more than 5 days from the onset of the symptoms). In COVID-19 patients' consolidation gradually increase over time. Bilateral, peripheral, middle/lower locations are the most frequent location (Vancheri et al., 2020). In some hospitals, the CXR examination is replaced or accompanied by CT scan, which showed a sensitivity of 97% for COVID-19 diagnosis (Ai et al., 2020), albeit with a limited specificity of 25%. Both CXR and CT have specific pros and cons, but the latter poses several logistic issues, such as the lack of availability of machines' slots, the difficulty of moving bedridden patients, and the long sanitization times. Furthermore, patients follow-up with CXR is simplified because it can be acquired at the patient's bed and, when required, directly at home (Zanardo et al., 2020).

Recently, artificial intelligence (AI) has been widely adopted to analyse CXR for several purposes, such as tuberculosis detection (Liu et al., 2017), abnormality classification and image annotation (Yan et al., 2019), pneumonia screening in pediatric and non pediatric patients (Radiological Society of North America, 2018), edema and fibrosis (Xu et al., 2018). Obviously, the challenge of COVID-19 pandemic has boosted the research efforts of AI in medical imaging and, according to the work presented by Greenspan et al. (2020), such applications may have an impact along three main directions, namely, detection and diagnosis, patient management and predictive modelling.

Regarding detection and diagnosis, AI is mainly used to detect the presence of COVID-19 patterns by processing CXR and/or CT images with deep neural networks (DNNs), such as convolutional neural networks (CNNs). DNNs were also applied to lesions segmentation or to produce a coarse localization map of the important regions in the image. For instance, Zhang et al. (2020) analysed CT scans collected from 4695 patients to differentiate novel coronavirus pneumonia from other types of pneumonia (bacterial, viral and mycoplasma pneumonia) and from healthy subjects. The classification was based on the combination of the segmented lung-lesion map and the normalized CT volumes. Experimental tests were performed on 260 patients, achieving an overall accuracy equal to 92.49%. Minaee et al. (2020) analysed 5000 chest X-ray images from publicly available datasets using four well known convolutional neural networks: ResNet-18, ResNet-50, SqueezeNet, and DenseNet-121. Two thousand images were used for training, whilst the models were tested on the other 3000, attaining a sensitivity rate equal to 98%, and a specificity rate of around 90% in detecting COVID-19 patients from their CXR.

The development of systems supporting patient management during hospitalization is mainly concerned with the monitoring of disease evolution in time. For instance, Gozes et al. (2020) proposed an image-based tool supporting the measurement of disease extent within the lungs. This severity biomarker is intended to help physicians in the decision-making process by tracking the disease severity over time.

Finally, predictive modelling mainly concerns with the development of models able to predict the progression of the disease. These approaches usually make use of both imaging and clinical data to predict the severity of the infection or the progression time, i.e. the time from the initial hospital admission to severe or critical illness, defined by death or the need for mechanical ventilation or the need for being transferred to the intensive care unit (ICU) (Zhang et al., 2020). Few applications have been recently developed within this category. For example, Greenspan et al. (2020) in their position paper presented preliminary and unconsolidated results on predicting the probability for a patient to be admitted to the ICU by exploiting quantitative features extracted from the lung region of the CXR images, vital parameters, comorbidities, and other clinical parameters. These data fed a random forest, which attained an area under the ROC curve (AUC) equal to 0.83. A survey offered by Wynants et al. (2020), compared 16 papers presenting prognostic models (8 for mortality, 5 for progression to severe/critical state and 3 for length of stay), and the AUC ranged from 0.85 up to 0.99. Nine of such papers used only clinical data for the analysis, whilst the others used clinical data and features extracted from CT images. The authors also argued that all 16 papers have a high risk of bias (Moons et al., 2019) due to the high probability of model overfitting and unclear reporting on intended use of the models. Still using CT images, two multicentric studies have been recently presented by Yue et al. (2020) and Chassagnon et al. (2020). The former included a cohort of 52 patients from five hospitals to predict short- or long-term hospital stays in patients with COVID-19 pneumonia. First, the CT scans were semi-automatically segmented and then for each lesion patch the authors extracted 1218 features, accounting for first-order, shape, second-order and wavelet measures. Second, a logistic model and a random forest were trained on the data from four hospitals, being tested on patients belonging to the fifth clinic. They attained balanced accuracies equal to 0.94% and 0.87%, respectively. The work presented by Chassagnon et al. (2020) aims to predict patient outcomes (severe vs. non-severe) prior to mechanical ventilation support and to suggest a possible prognosis within three available groups (short-term deceased, long-term deceased, long-term recovered). To these goals, they searched for a subset of discriminative features from several image texture descriptors computed from CT scans and a few clinical data (i.e. age, gender, high blood pressure, diabetes, body mass index). On a cohort of 693 patients, an ensemble of classifiers separated patients with severe vs. non-severe outcomes and it correctly identified the prognosis with balanced accuracies equal to 70% and 71%, respectively.

This analysis of the literature shows that the development of AI-based models predicting the outcomes of COVID-19 patients still deserves further research efforts. On the one side, in the context of COVID-19 prognosis, except for the very preliminary results anticipated by Greenspan et al. (2020), all the works in literature used CT scans although CXRs are considered a viable alternative by the American College of Radiology (2020). On the other side, sharing patient data from studies as well as creating new data sets collected in clinical practice is fundamental for the AI com-

[1] https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases

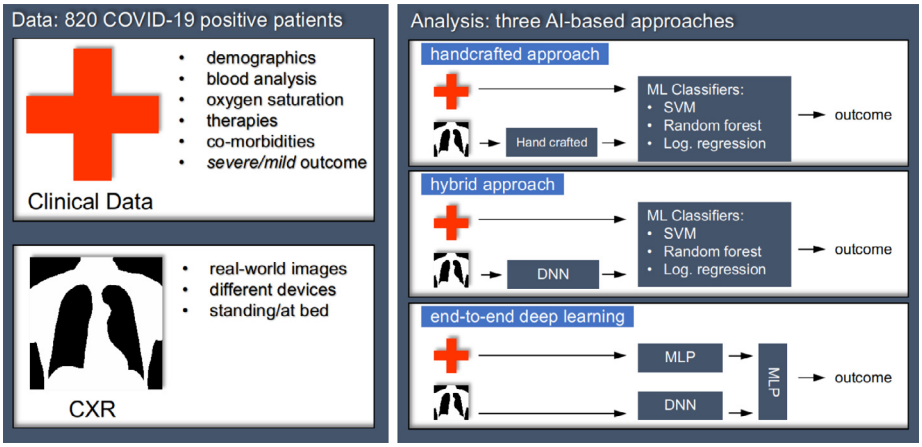


Fig. 1. Overview of the method for automatic prognosis of COVID-19 in two classes, namely mild and severe. Our works includes data collected in 6 independent cohorts, resulting in 820 COVID-19 patients. For each, we collected several clinical attributes, combined with quantitative imaging biomarkers computed by handcrafted features or automatically computed by CNNs.

munity (Leeuwenberg and Schuit, 2020), since many researchers do not have the possibility of collecting clinical data and images from different clinical centres. To address both concerns, this work investigates three AI-based approaches to predict clinical outcome integrating clinical and imaging data. Indeed, in addition to clinical information consisting of general information, laboratory data and comorbidities, such approaches use quantitative information extracted from the CXR images, which are also referred to as image features or quantitative biomarkers in the following. The first approach computes handcrafted texture features to be used by a common classifier, the second approach automatically extracts image descriptors by using a CNN, while the third approach is fully based on DNNs, processing both clinical and image data (Fig. 1). As a further contribution, this work introduces a novel dataset including clinical data and CXR images from 820 patients with COVID-19 who were hospitalized in six hospitals in Italy. To each patient we associated prognostic information related to the clinical outcome. Our work therefore offers also a first quantitative analysis of this new repository that can be used by other researchers and practitioners as a baseline reference.

In synthesis, the main objectives of this work are:

1. to present an evaluation of three state-of-the-art learning approaches to predict future severe cases at the time of hospitalization, which are specifically designed to use either handcrafted or learned image features, together with clinical data;
2. to boost the research on AI-based prognostic models to support healthcare systems in the fight against COVID-19 pandemic by making publicly available a repository of CXR images and clinical data collected in a true environment during the first wave of the pandemic emergency, which include common real-world issues such as missing data, outliers, different imaging devices, poorly standardized data. The repository would also facilitate external validation of learning models developed in this field.

The rest of this manuscript is organized as follows: next section describes the dataset we collected and that we are making publicly available. Section 3 introduces the methodology we adopted, whilst Section 4 presents the classification results achieved. Section 5 discusses our findings providing also concluding remarks.

2. The AIforCOVID dataset

This study includes the images and clinical data collected in six Italian hospitals at the time of hospitalization of symptomatic patients with COVID-19, during the first wave of emergency in the country (March–June 2020). Such data was generated during the clinical activity with the primary purpose of managing COVID-19 patients within the daily practice and they were retrospectively reviewed and collected, after patients' anonymization. Ethics Committee approval was obtained (Trial-ID: 1507; Approval date: April 7th, 2020) and all data were managed in accordance with the GDPR regulation. Furthermore, we randomly assigned to each centre a symbolic label, from A up to F.

The 820 CXR examinations reviewed in this study were performed in COVID-19-positive adult patients at the time of hospital admission (Table 1): all the patients resulted positive for SARS-CoV-2 infection at the RT-PCR test (Yang et al., 2020). In 5% of such cases, the positivity to the swab was obtained only at the second RT-PCR examination. In the different centres, CXR examinations were performed using different analog and digital units, and the execution parameters were settled according to the patient conditions. Paired with CXR examinations, we collected also relevant clinical parameters listed in Table 2.

According to the clinical outcome, each patient was assigned to the *mild* or the *severe* group. The former contains the patients sent back to domiciliary isolation or hospitalized without ventilatory support, whereas the latter is composed of patients who required non-invasive ventilation support, intensive care unit (ICU) and deceased patients. Fig. 2 shows four difficult examples of CXR images within the dataset: indeed, panels A and B show two images of patients with severe outcome whilst the radiological visual inspection may suggest severe and mild prognoses, respectively. Similarly, panels C and D show two images of patients with mild outcome whilst a radiologist may report severe and mild prognosis, respectively.

During an initial data quality cleaning, we double-checked with the clinical partners the anomalous data and the outliers, i.e. those values lying outside the expected clinical range or identified applying the interquartile range method, which were then corrected when needed. Categorical variables values were homogenized to a coherent coding, such as 0 and 1 values for binary variables like comorbidities and sex, and we adopted the string "NaN" to denote missing data. No exclusion rule was applied for images based on device type or brand (e.g. digital or analog devices) or patient positions (standing or at bed), whereas X-ray images taken with lateral projection were excluded because they were not available for patients whose images were acquired in the lying position. In the case of multiple CXR images delivered for the same patient, the dataset contains only the first one. It is worth noting that the presence of missing entries in the clinical data mostly depends upon

Table 1
Patient distribution across the hospitals where the data were collected.

Hospital	Number of patients	Mild class prior probability	Anterior posterior (AP) projection prior probability
A	120	29.2%	81.67%
B	104	56.7%	97.12%
C	31	25.8%	90.32%
D	139	54.7%	38.85%
E	101	54.5%	87.13%
F	325	46.5%	98.46%
Overall	**820**	**46.8%**	**83.72%**

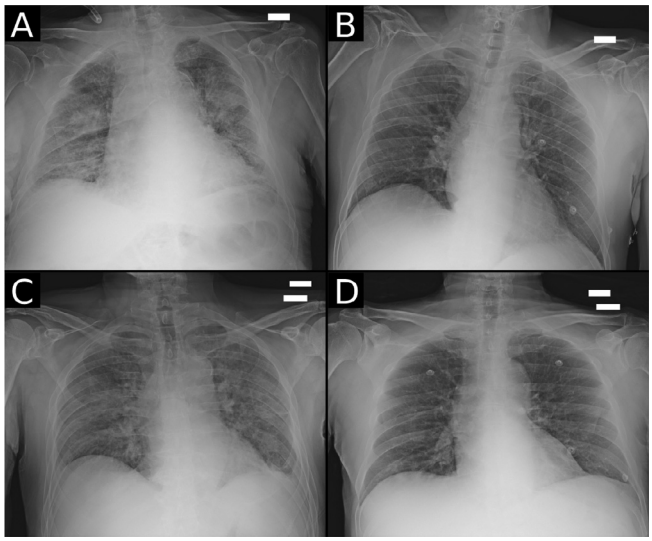


Fig. 2. Examples of CRX images of patients with COVID-19 available within the dataset. Panels A and B show two images of patients with severe outcome whilst the radiological visual inspection may suggest severe and mild prognoses, respectively. Similarly, panels C and D show two images of patients with mild outcome whilst a radiologist may suggest severe and mild prognosis, respectively, based on the visual interpretation.

the procedures carried out in the individual hospitals as well as upon the pressure due to the overwhelming number of patients hospitalized during the COVID-19 emergency. For the sake of completeness, the rate of missing data is reported in the last column of Table 2.

CXR images were collected in DICOM format and, for anonymization constrains, all the fields but a set of selected metadata related to acquisition parameters were blanked in the DICOM header (e.g. image modality, allocated bits, pixel spacing, etc.).

All the images in the repository are currently stored using 16 bits, while acquisition precision varies: 13.5% were acquired at 10 bits precision, 35.4% at 12 bits, 46.6% at 14 bits and 4.5% using the full 16 bits precision. Furthermore, all the images were acquired with isotropic pixel spacing ranging from 0.1 mm to 0.2 mm. The most common pixel spacing is 0.15 mm, 0.1 mm and 0.16 mm for 43.9%, 13.7% and 13.6% of images respectively. Image sizes, in pixels, are distributed as follows: 33.4% of the images have 2336×2836 pixels, 13.5% of images have 3520×4280 pixels and 10.1% of the images have 3480×4240 pixels. The other images have a number of rows ranging from 1396 up to 4280, whilst the number of columns ranges from 1676 up to 4280.

2.1. Statistical analysis of clinical data

We performed a statistical analysis applying the Mann–Whitney *U* test to compare mild- and severe-groups in case of continuous variables, whereas we used the z test with Yates continuity correction for analysing proportions.

Summary statistics are reported in Table 2. For continuous variables median and interquartile range (IQR) were reported. For categorical variables we reported patients' proportions expressed as percentage. For statistical analysis a p-value lower than 0.05 was considered significant.

The analysis evidenced that females represented the 32% ($n = 266$) of the total population and they were significantly ($p < 0.001$) older (median age 70 years, IQR 57–80 years) than males (median age 64 years, IQR 53–74 years). Furthermore, 522 out of 820 (63%) patients had at least one comorbidity (Fig. 3).

In agreement with widely reported demographic data showing that older patients had more severe outcome, in our dataset, the patients of severe-group (70 years, IQR 60–79) were significantly ($p < 0.001$) older than those belonging to the mild-group (60 years, IQR 49–72 years). Three hundred twenty patients out of 436 (73%) of the severe group had at least one comorbidity; in the mild-group they were 194 out of 384 (51%).

Moreover, 47% (384/820) of patients belonged to the mild-group, of which 157 (41%) were females and significantly ($p = 0.015$) older (median 63 years, IQR 50–76 years) than males (median 59 years, IQR 48–69 years). In the severe-group were 436 out of 820 (53%) patients, of which 109 (25%) were females and significantly ($p < 0.001$) older (median age 78 years, IQR 67–85 years) than males (median age 67 years, IQR 57–76 years). Severe group consisted of 43% (189/436) of patients hospitalized with non-invasive ventilation support, 18% (79/436) of patients in ICU, and 38% (168/436) of dead patients. Regarding the dead patients' subgroup, the mean age was 78 (IQR 68–84) years: the youngest was 43yo while the oldest was 97yo. Among dead patients, 97% of them had at least one co-morbidity, while 17% had five comorbidities reported. The majority (72%, 121 out 168) of dead patients were male.

3. Methods

We investigated three AI-based prognostic approaches covering well-known methodologies with the intent to offer to researchers and practitioners a reference baseline to process the data available within the AIforCovid dataset. Furthermore, for the sake of an easy and fair comparison and to foster further research in this field, we detail also the adopted validation procedures, recommending others to measure models performance at least as reported here.

As schematically depicted in Fig. 1, the first learning approach employs first order and texture features computed from the images, which are mined together with the clinical data feeding a supervised learner. In the following, it is shortly referred to as *handcrafted approach*, and it is presented in Section 3.6.

In the last decade, we have assisted to the rise of deep artificial neural networks, which have attained outstanding performance in many fields. Recently, DNNs such as convolutional neural networks have been applied also to COVID-19 imaging mostly for diagnostic purposes (Greenspan et al., 2020). On this basis, the second approach presented here mixes automatic features computed by a CNN with the clinical data. Shortly, we used a pre-trained CNN as a

Table 2
Description of the clinical data available within the repository. First and second columns report variables label and description. Summary statistics for the overall population and for the two patients groups are reported in the following columns. For continuous variables median and interquartile range are reported, for categorical variables proportions are reported. Feature names followed by '+' were not used for the analysis described in this work. *P*-values lower than 0.05 were considered significant. * Mann–Whitney *U* test. † *z*-test for proportions with Yates continuity correction. ‡ Fisher exact test.

Name	Description	Overall-population	Mild-group (A)	Severe-group (B)	A vs. B *p*-value	Missing data (%)
Active cancer in the last 5 years	Patient had active cancer in the last 5 years (% reported)	7%	5%	8%	<0.05†	1.4
Age	Patient's age (years)	64; 54–77	60; 49–72	70; 60–79	<0.001*	0
Atrial Fibrillation	Patient had atrial fibrillaton (% reported)	9%	5%	11%	<0.01†	2.2
Body temperature (°C)	Patients temperature at admission (in °C)	38; 37 and 38	38; 37 and 38	38; 37 and 38	0.171	8.8
Cardiovascular Disease	Patient had cardiovascular diseases (% reported)	35%	23%	40%	<0.001†	1.7
Chronic Kidney disease	Patient had chronic kidney disease (% reported)	6%	4%	9%	<0.01†	1.4
COPD	Chronic obstructive pulmonary disease (% reported)	7%	4%	10%	<0.01†	1.4
Cough	Cought (%yes)	54%	59%	50%	<0.05†	0.5
CRP	C-reactive protein concentration (mg/dL)	57; 24–119	42; 17–75	103; 48–163	<0.001*	3.5
Days Fever	Days of fever up to admission (days)	3; 2–4	3; 2–4	3; 2 and 3	0.289	10.96
D-dimer	D-dimer amount in blood	632; 352–1287	549; 262–909	820; 438–2056	<0.001*	77.6
Death+	Death of patient occurred during hospitalization for any cause	168	0	168	–	–
Dementia	Patient had dementia (% reported)	4%	3%	6%	0.087	1.8
Diabetes	Patient had diabetes (% reported)	16%	10%	21%	<0.001†	1.4
Dyspnea	Patient had intense tightening in the chest, air hunger, difficulty breathing, breathlessness or a feeling of suffocation (%yes)	50%	37%	62%	<0.001†	0.4
Fibrinogen	Fibrinogen concentration in blood (mg/dL)	607; 513–700	550; 473–658	615; 549–700	<0.001*	73.6
Glucose	Glucose concentration in blood (mg/dL)	110; 96–130	104; 93–121	114; 101–139	<0.001*	20.6
Heart Failure	Patient had heart failure (% reported)	2%	1%	3%	0.157	2.3
Hypertension	Patient had high blood pressure (% reported)	46%	38%	54%	<0.001†	1.4
INR	International Normalized Ratio	1.13; 1.07–1.25	1.11; 1.06–1.20	1.15; 1.08–1.28	0.004*	28.8
Ischemic Heart Disease	Patient had ischemic heart disease (% reported)	15%	11%	18%	<0.01†	18.3
LDH	Lactate dehydrogenase concentration in blood (U/L)	320; 249–431	271; 214–323	405; 310–527	<0.001*	24.6
O_2 (%)	Oxygen percentage in blood (%)	95; 90–97	96; 94–98	92; 87–96	<0.001*	16.5
Obesity	Patient had obesity (% reported)	9%	6%	11%	0.058	36.1
$PaCO_2$	Partial pressure of carbon dioxide in arterial blood (mmHg)	33; 30–36	34; 30–37	33; 30–35	0.116	15.4
PaO_2	Partial pressure of oxygen in arterial blood (mmHg)	69; 59–80	73; 67–81	64; 54–76	<0.001*	15.3
PCT	Platelet count (ng/mL)	0.19; 0.09–0.56	0.09; 0.05–0.26	0.28; 0.13–0.72	<0.001*	71.8
pH	Blood pH	7; 7 – 7	7; 7 – 7	7; 7 – 7	<0.001*	17.3
Position+	Patient position during chest X-ray (%supine)	78%	68%	87%	<0.001†	0
Positivity at admission	Positivity to the SARS-CoV-2 swab at the admission time (% positive)	95%	94%	96%	0.142	4.7
Prognosis	Patient outcome, see Section 2 (% cases)	–	46.8%	53.2%	0.468†	0.0
RBC	Red blood cells count (10^9/L)	4.65; 4.26–5.07	4.70; 4.34–5.11	4.59; 4.13–5.03	<0.001*	3.0
Respiratory Failure	Patient had respiratory failure (% reported)	1%	100%	2%	0.131	19.0
SaO_2	arterial oxygen saturation (%)	95; 91–97	96; 94–98	92; 87–96	<0.001*	59.2
Sex	Patient's sex (%males)	68%	59%	75%	<0.001†	0
Stroke	Patient had stroke (% reported)	4%	3%	4%	0.447	2.3
Therapy Anakinra+	Patient was treated with Anakinra (%yes)	100%	0%	0%	–	10.8
Therapy anti-inflammatory+	Patient was treated with anti-inflammatory drugs therapy (%yes)	55%	53%	57%	0.243	13.5
Therapy antiviral+	Patient was treated with antiviral drugs (%yes)	47%	44%	50%	0.129	10.7
Therapy Eparine +	Patient was treated with eparine (no; yes; prophylactic treatment; therapeutic treatment)	56.6%; 11.5%; 28%; 3.9%	73.3%; 8.3%; 17.2%; 1.1%	39.9%; 14.7%; 38.8%; 6.6%	<0.001‡	13.4
Therapy hydroxychloroquine +	Patient was treated with hydroxychloroquine (%yes)	59%	56%	62%	0.118	11.6
Therapy Tocilizumab +	Patient was treated with Tocilizumab (%yes)	9%	2%	15%	<0.001†	12.4
WBC	White blood cells count (10^9/L)	6.30; 4.73–8.42	5.58; 4.32–7.17	7.10; 5.25–9.80	0.012	0.7

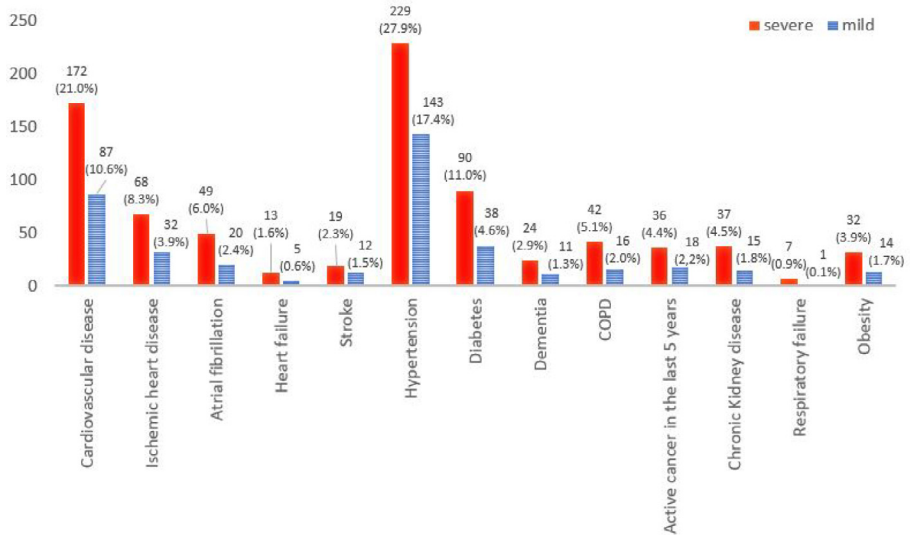


Fig. 3. Comorbidity distributions between groups. For all data, value and percentage referred to the total population was indicated.

Table 3
Summary of the operations common to the three AI approaches.

Method	Operations			
	Data imputation	Image standardization	Lung segmentation	Feature selection
Handcrafted	✓	✓	✓	✓
Hybrid	✓	✓	✓	✓
End-to-end DL	✓	✓		

CXR feature extractor. The output of the last fully-connected layer was then provided as input for a SVM classifier, together with the clinical features.

In the following, it is shortly referred to as *hybrid approach*, and it is presented in Section 3.7.

The third approach exploits together the clinical data and the raw CXR using a multi-input convolutional network to predict patients' outcome. In order to handle data from such different sources, the network consists of two dedicated input branches, while higher-level features from both sources are concatenated in the last layers before the classification output. In the following, this approach is shortly referred to as *end-to-end deep learning approach*, and it is detailed in Section 3.8.

Note that all such approaches do not use the therapy-related variables included in the dataset because, albeit therapy could influence the final outcome, it is also dependent on the outcome (i.e. patients who required intensive care were administered with specific therapies). Furthermore, the classification task defined considers only the data collected at the time of hospitalization and, therefore, in a true clinical scenario, information on the administered therapy would not be available. For this reason, the use of those variables could be misleading.

Before presenting in detail each of the three approaches, following Sections 3.1 and 3.2 describe data imputation and image standardization. Furthermore, Section 3.3 presents the framework used to segment the lung, whereas Section 3.4 describes the feature selection approach and the classifiers adopted, which are the same across the three methods to facilitate their comparison. Table 3 summarize the common operations applied by each of the three AI methods and Section 3.5 introduces the procedure adopted to validate the learning models.

3.1. Data imputation

To deal with missing data, univariate data imputation estimates missing entries by using the mean of each column in which the missing values are located. We preferred this approach to multivariate or prediction-based imputation methods since it is known to work well when the data size is not very large, and it can prevent data loss which results from brute force rows and columns removal. Furthermore, preliminary results not shown here confirmed such observations. As reported in the second column of Table 3, imputation was performed before each learning paradigm worked on the data.

3.2. Images standardization

CXR images collected for this study were acquired with different devices and acquisition conditions, as mentioned in Section 2. For this reason, we applied image normalization that, to a large extent, is the same for all the three methods. Indeed, for the handcrafted approach pixels values were normalized to have zero mean and unit standard deviation, whilst the images were resized to 1024×1024 pixels using bilinear interpolation. For the hybrid approach, a segmentation network was used to identify the square box containing the lungs, in a way to crop only the region of interest, as detailed in the next section. The images were then normalized and resized to a dimension equal to 224×224 pixels, as we employed early processing layers pre-trained on the ImageNet dataset, which consists of images of this size. Similarly, in the end-to-end DL approach images were resized to 224×224 pixels, without prior cropping, and normalized as in the previous cases.

3.3. Lung segmentation

When needed, to segment the lung we apply a semi-automatic approach that initially delineates the lung borders using a U-Net, which is a convolutional neural network architecture for fast and precise segmentation of images. In this respect, it is well known that the semantic segmentation provided by this deep network has proven to have very satisfactory performance when using medical

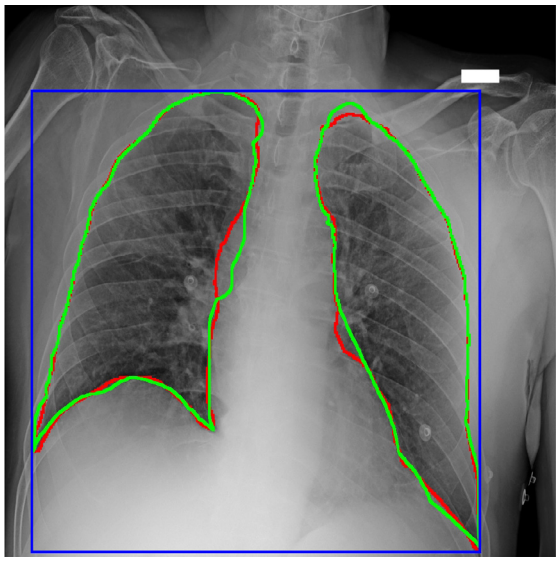


Fig. 4. Example of the lung segmentation results. Green line: manual segmentation, red line: segmentation returned by the U-Net, blue line: bounding box from U-Net segmentation. For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.

images (Zhou et al., 2018; Hesamian et al., 2019; Rajaraman et al., 2020a).

The network adopted here was already trained on non-COVID-19 lung CXR datasets,[2] namely the Montgomery County CXR set (MC) (Jaeger et al., 2014) and the Japanese Society of Radiological Technology (JSRT) repository (Shiraishi et al., 2000), using an Adam optimizer and with a binary cross-entropy loss function. Furthermore, during training, a random augmentation phase composed of rotation ($\pm 10^{\circ}$), horizontal and vertical shift (± 25 pixels), and zoom (0–0.2) was applied. Furthermore, the batch size was set to 8 and the number of epochs was equal to 100. The MC dataset contains 7470 CXR collected by the National Library of Medicine within the Open-i service, whereas the JSRT repository is composed of 247 CXR with and without a lung nodule. The U-net requires input images represented as 3-channel 256×256 matrices and, hence, grayscale images were copied to all the channels and then resized. Furthermore, we normalized the pixel intensities as detailed in Section 3.2. After these transformations, each image was passed through the convolutional network and all the pixels were classified as foreground (i.e. the lung) or as background. To check if the network worked well, all the images in the repository were segmented by two expert radiologists working in parallel using a consensus strategy (Fig. 4), permitting us to assess the U-Net segmentation performance. We found that the network provides a Jaccard index and a Dice score equal to 0.896 and 0.942, respectively. We deem that such performance are satisfactory since it is only needed to recover the bounding box, as in the hybrid approach presented below, while it would not be sufficient for exact lung delineation needed by the following handcrafted approach.

3.4. Feature selection and classifiers

In general, we had a large number of descriptors that suggested us to apply a feature selection stage, which was set up in two steps. The first is a coarse step that runs a univariate filtering based on mutual information as a score function to pre-select a reduced set of image descriptors, whatever the approach used for their computation. The calculation of mutual information between continuous features with the discrete class variable was addressed by estimating the entropy from the k-nearest neighbours distances (Ross, 2014).

The second feature selection step merges the pre-selected imaging features with the clinical data. To this end, we applied a wrapper approach, namely the Recursive Feature Elimination and Cross-Validated selection (RFECV) method (Guyon et al., 2002), which receives as input the pre-selected imaging descriptors and the 34 clinical features. Indeed, the RFECV is fed by an increasing number of pre-selected imaging descriptors (D_{pr}): fine-grained sampling was carried out for $D_{pr} \leq 10$ applying a step of 2; for $10 < D_{pr} \leq 50$, D_{pr} was sampled with step of 5; finally, RFECV was fed with all the image features. RFECV applies a pruning procedure that starts considering all features in the dataset and recursively eliminates the less important according to a feature's importance score calculated using a classifier. Note that the optimal number of features is selected by RFECV using nested 5-fold cross-validation on the training set. With reference to the base learner we evaluated three different computational paradigms: Logistic Regression (LGR); Support Vector Machines with a linear kernel (SVM); and Random Forests (RF). For all parameters in the adopted models we used the default values provided by the libraries, without any fine tuning. Indeed, we were not interested in the best absolute performance. Moreover, Arcuri and Fraser (2013) empirically observed that in many cases the use of tuned parameters cannot significantly outperform the default values of a classifier suggested in the literature.

3.5. Models validation

Model validation for the three tested methods consists of k-fold and leave-one-centre-out cross validation. For each cross-validation run, the training fold was used for data normalization, parameters' estimation and/or features' selection depending on the applied method. Classification performance assessment was carried out using testing fold data only; k-fold cross-validation was repeated with k equal to 3 and 10 with 20 repetitions. In leave-one-centre-out (LOCO) cross validation, in each run the test set is composed of all the samples belonging to one centre only, while the others were assigned to the training set. When needed, the validation set was extracted from the training set using any policies (such as random selection, hold-out, nested cross validation, etc.), and considering also the computational burden.

Performance of the learning models was measured in terms of accuracy, sensitivity and specificity, reporting the average and standard deviation of each experiment. When needed, we ran the pairwise two-sided Mann Whitney U test to compare the results provided by two methods, whereas we performed the Kruskal–Wallis test followed by the Dunn's test with Bonferroni correction for multiple comparisons. In the rest of the manuscript we assume that the pairwise two-sided Mann Whitney U test was performed by default, otherwise we will specify the test used.

3.6. Handcrafted approach

The handcrafted approach first computes parametric maps of the lungs segmented in the CXR image; second it extracts several features that are then provided together with the clinical data to a supervised learner.

To segment the lung we applied the approach presented in Section 3.3 but, as mentioned there, we deem that the segmentation performance is not satisfactory for exact lung delineation. For this reason, the lung masks are then reviewed by expert radiologists and then used to compute the handcrafted features as follows.

[2] The network is available as detailed in the reference denoted as Imlab-UIIP (2020).

From the segmented lungs we computed the parametric maps using a pixel-based approach as proposed by Penny et al. (2011). Pixels values of the parametric maps were obtained by computing first- and second-order radiomic features on a 21×21 sliding window running over each pixel of the entire lung region. First-order measures describe the statistical distribution of tissue density inside the kernel; from its grey levels' histogram, we extracted 18 descriptors, whose formal presentation is offered in the dedicated section of the supplementary file. Second-order descriptors are based on the Grey Level Co-occurence Matrix (GLCM): at each location, we got a GLCM image, where we computed 24 Haralick descriptors (Haralick et al., 1973) detailed in the same section of the supplementary file, as before. This procedure returned 42 parametric images (18 First-order + 24 GLCM) for each CXR image, where we finally computed seven statistics, namely: mean, median, variance, skewness, kurtosis, energy and entropy. This resulted in 294 image features (i.e. 7 statistics by 42 parametric maps).

To cope with the large number of descriptors we proceeded as described in Section 3.4, adopting the base learners already described there. Then, for each tested classifier, given the set and number of descriptors selected by the wrapper approach in the nested cross-validation fashion, we trained the same classifier on the whole training fold and measured recognition performance on the test fold.

3.7. Hybrid approach

The hybrid approach integrated the output of a pre-trained deep network and the set of clinical measures. The pipeline worked as follows: first, we applied a pre-trained deep neural network to segment the lungs; second, a convolutional neural network was trained to extract relevant features from the CXR images; third, we concatenated the deep features with the clinical ones; fourth, we performed a feature selection step as reported in Section 3.4; fifth, we trained a supervised classifier to accomplish the binary classification task. In the following we will illustrate these steps.

As mentioned before, the image repository is composed of CXR images collected in multiple hospitals, using different machines with different acquisition parameters. This resulted in a certain degree of variability among the images, where the lungs have also different sizes. To cope with this issue, we adopted the segmentation net already discussed in Section 3.3, which boosts the performance of the feature extraction network by locating the lungs. Differently from before, where the U-Net was used to pre-segment the lungs whose borders were manually refined, here we adopted a fully automated approach since the segmentation mask given by the network was used to extract the rectangular bounding box containing the ROI. Now there was no need for any manual intervention since the performance at the level of ROI bounding box segmentation was satisfactory, when compared with human's annotation. Indeed, the Jaccard index and the Dice score were now equal to 0.929 and 0.960, respectively.[3]

Next, each ROI was resized to a square so that the longest side of the ROI was mapped to the square side, and the other ROI side was resized accordingly. Each cropped image was then passed to a deep neural network to extract the features, where we performed a transfer learning process as follows. Indeed, preliminary experiments showed that such an approach gave better results than starting the training from scratch.

Furthermore, to alleviate the risk of overfitting and reduced generalization typical of learning models working with medical images, we pre-trained several state-of-the-art network architectures previously initialized on other repositories. To this end, we used the chest X-ray images dataset presented by Mooney (2017); Kermany et al. (2018), which consists of 5863 CXR images classified as pneumonia or normal by two expert physicians. This would allow the networks to learn modality-specific feature representations (Rajaraman et al., 2020a; 2020b). After this step, such models were fine-tuned on our image dataset. In a first stage we tested in 10-fold cross validation these networks: Alexnet (Krizhevsky, 2014, VGG-11, VGG-11 BN, VGG-13, VGG-13 BN, VGG-16, VGG-16 BN, VGG-19, VGG-19 BN Simonyan and Zisserman, 2014), ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-152 (He et al., 2016), ResNext (Xie et al., 2017), Wide ResNet-50 v2 (Zagoruyko and Komodakis, 2016), SqueezeNet-1.0, SqueezeNet-1.1 (Iandola et al., 2016), DenseNet-121, DenseNet-169, DenseNet-161, DenseNet-201 (Huang et al., 2017), GoogleNet (Szegedy et al., 2015), ShuffleNet v2 (Ma et al., 2018) and MobileNet v2 (Sandler et al., 2018). Then, to reduce the computational burden, the top-five networks (i.e. VGG-11, VGG-19, ResNet-18, Wide ResNet-50 v2, and GoogleNet) underwent all the experiments described in Section 3.5. In all the cases, we changed the output layer of the CNNs, using two neurons, one for each class. Moreover, image standardization as described in Section 3.2 was performed. We also augmented the training data by independently applying the following transformations with a probability equal to 30%: vertical and horizontal shift $(-7, +7)$, y-axis flip, rotation $(-175^\circ, +175^\circ)$ and elastic deformation ($\sigma = 7$, $\alpha = [20, 40]$). Training parameters were: a batch size of 32 with a cross-entropy loss, a SGD optimizer with learning rate of 0.001 and momentum of 0.9, with max epochs sets equal to 300 and an early stopping criterion fixed at 25 epochs, using the accuracy on the validation set. In this respect, it is worth noting that we also performed a preliminary optimization of CNN hyperparameters using Bayesian Optimization (Mockus, 2012), and we found that the results did not statistically differ from those achieved using the aforementioned values, according to the Wilcoxon's test (with $p = 0.05$). Furthermore, this finding agrees also with what reported by Arcuri and Fraser (2013), already summarized at the end of Section 3.6.

Once the deep networks were trained, we integrated the automatic features they computed with the clinical information. To this goal, we extracted the last fully connected layer for each network, which was used as a vector of features for each patient; accordingly, on the basis of the network we were using, the number of automatic features varied between 512 and 4096 (i.e. it is 512 for ResNet-18, 1024 for GoogleNet, 2048 for Wide-ResNet-50 v2, and 4096 for VGG-11 and VGG-19) Each of such sets of automatically computed descriptors was combined with the clinical data and, to avoid to overwhelm the latter, the number of features in the former was reduced by a coarse selection stage using the univariate approach already described in Section 3.4. Furthermore, we then applied the same wrapper approach to investigate if the combination of automatic and clinical features had a degree of redundancy. Straightforwardly, to avoid any bias all the operations described so far were performed respecting the training, validation and test split introduced before, and ensuring that the test was not used in any stage except for the final validation.

Finally, the selected features were used to classify each patient in the two classes already mentioned, i.e. mild and severe, as reported in the last part of previous subsection, and using the same learners already mentioned.

3.8. End-to-end deep learning (DL) approach

The end-to-end DL approach was designed so that clinical information could influence the generation of useful features in image classification and vice-versa. Two different variants have been

[3] In only one case the segmentation network did not segment the lungs; in this case, the entire original image is used.

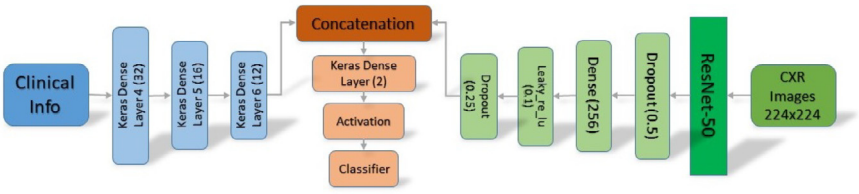


Fig. 5. Workflow of the end-to-end deep learning approach.

tested: in the first, CXR images were modified with the addition of an extra layer, in which pixels in fixed positions would code properly normalized clinical information, while the remaining pixels were filled with uniform white noise. The second variant consists in a multi-input network, which received separately CXR images and clinical information. Eventually, the latter variant proved to perform slightly better and it will be described in the following.

The architecture we adopted is composed of three main sections: one branch for each input accepts raw data and processes them to obtain a small number of relevant features, while a final common path concatenates the output features of the previous branches and uses them to provide the actual classification. A representation of the network can be found in Fig. 5.

Several different image classification networks have been tested (VGGs, ResNet, Inception and Xception variants[4]), with the ResNet-50 architecture resulting either the best performer or tied for best performer in all considered conditions (see Supplementary Tables 5 and 6).

The network has been adopted up to the last convolutional layer, while the final fully-connected layer and classification section have been removed. The number of generated output features has been reduced by a dropout layer with probability of 0.5, followed by a 256-neurons fully-connected layer, a leaky ReLU and a final dropout layer with probability 0.25.

All network tested were pre-trained on the ImageNet dataset, then on the same publicly available repository of CXR images introduced in Section 3.7 with the task of discriminating between healthy subjects and pneumonia patients (Mooney, 2017); finally, the network was trained on the dataset presented here. We found that, with our architecture, data augmentation did not improve final classification performance and was, therefore, excluded from the processing pipeline. On the other hand, pre-training led to more consistent results, as discussed in Section 5.

The clinical information branch is a multi-layer perceptron (MLP): it is composed of a sequence of alternating fully-connected and non-linear layers; the adopted architecture consists of three fully-connected layers of decreasing size (32, 16 and 12 neurons), alternating with Rectified Linear Units (ReLUs).

The common section of the network consists of a concatenation layer, which receives a total of 268 inputs (256 from the image branch and 12 from the clinical information branch) and feeds them to the actual classification section of the network (2-neurons fully-connected, softmax and classification layers).

In order to evaluate the impact of each data source, the model was trained as described above, as well as in two different versions modified to accept one data source only (i.e. changes consists in removal of one input branch and concatenation layer and with a change in the number of neurons in the final fully-connected layer). All versions underwent the same training procedure: a 20-epochs training phase with a SGD optimizer with a momentum set to 0.9. The weights used on the test set correspond with the iteration resulting in the lowest loss. The learning rate is fixed and equal to 10^{-4}, while the batch size is set to 16.

[4] Please note that when using the Inception and Xception variants the input images were resized to 299×299 rather than 224×224, as already described in Section 3.2.

4. Results

This section reports the results attained using the three approaches mentioned so far in staging the patients with COVID-19 in severe and mild classes. The goal is to provide a baseline characterization of the performance achieved integrating quantitative image data with clinical information by using state-of-the art approaches.

Tables 4 and 5 present the best recognition performance attained by each of the learning methods when the experiments were executed according to the 10-fold and LOCO cross validation, respectively (see Section 3.5 for further details). In the former case, the results are averaged over the 20 repetitions. Furthermore, for the sake of readability we omit to report the results achieved using the 3-fold cross validation since they are consistent with those performed in the 10-fold fashion. For the sake of completeness the interested readers can refer to the supplementary material to navigate all the results attained (Supplementary Tables 1, 3 and 5).

The first two rows in both tables report the performance in discriminating between patients with mild and severe prognosis attained using clinical data only. In this respect, the row denoted by Machine Learning (ML) shows the best performance achieved by the RFECV and by the learners described in the last part of Section 3.6, whereas the row denoted by Deep learning (DL) reports the performance returned by the multi-layer perceptron described in Section 3.8. In the case of experiments performed in 10-fold cross validation (Table 4), the best accuracy is up to 75.7%, it is attained by an SVM retaining on average 11 clinical features, and the sensitivity and the specificity are almost balanced. This latter observation can be expected since the a-priori class distribution is not skewed. We also notice that the use of a deep network is sub-optimal in the classification task based on clinical information alone: this is likely due to the fact that, in contrast with the image case, pre-training of the network was impossible, due to the custom nature of input data. As a consequence, it is possible that the available number of samples was not sufficient to train the network to optimal performance. The same observations hold also in the case of the experiments performed in a LOCO modality (Table 5), and it is worth noticing the performance drops for both the ML and DL approaches. This can be due to the variation of data distribution among the centres, limiting the generalization capability of the learners. Again, the readers can refer to the supplementary material to navigate all the results attained in LOCO cross-validation (Supplementary Tables 2, 4 and 6).

In both Tables 4 and 5, the next two sections report the performance attained by the three methods described in Section 3 using only the CXR images and merging together the images with the clinical data, respectively. With reference to the results reported in the section "CXR images", they show that the use of the images only does not achieve the same performance obtained using the clinical data, whatever the method applied (Table 4). Furthermore, the fact that the end-to-end DL has better results than the hy-

brid approach suggests that the fully connected portion of the CNN better exploits than a supervised classifier the information provided by the convolutional layers. In the case of the experiments performed in LOCO modality, there are still gaps with the results achieved using clinical data only, suggesting that all the learners suffer from the variability induced by the different centres. Turning our attention to the results shown in section "Clinical data and CXR images", in the case of the experiments performed in 10 fold cross validation we notice that the integration between the two sources of information provides some benefits, permitting in some cases to improve the classification performance. Indeed, the hybrid approach achieves an accuracy up to 76.9%, using the automatic features computed by the convolutional layers of the GoogleNet and an SVM classifier. The end-to-end DL approach slightly improves the performance with respect to the ones attained using only the images, suggesting that an approach fully based on DNN is not beneficial in this case, needing for further investigation. In the case of the experiments run in LOCO mode we found that the integration of clinical data and CXR images is beneficial as the largest accuracy is up to 75.2%, with improvements in terms of sensitivity and specificity.

Furthermore, to see if there exists a statistically significant difference between the various performance, we ran the Kruskal–Wallis and the Dunn test with Bonferroni correction for multiple comparisons ($p < 0.05$): the results are reported in the supplementary material (Supplementary Figs. 1, 2 and 3, which refer to the handcrafted, hybrid and end-to-end approaches, respectively). In the case of the handcrafted approach, this statistical analysis shows that in almost all the experiments the results achieved by the three learners (LGR, SVM and RF) are not different, at the given significance level. In the case of the hybrid approach (Supplementary Fig. 2), we find that each of the best learner reported in Tables 4 and 5 has performances that are statistically different by large part of the other learners. In the case of the end-to-end approach, Supplementary Fig. 3 shows that the ResNet50 provides performance statistically different from the other architectures in all the cases except one.

As a final point, the results mentioned so far were achieved using the following computational resources and deep learning frameworks. For the handcrafted approach we used Python-3.8.3, scikit-learn-0.23.1, pandas-1.0.5, numpy-base-1.18.5, and two NVIDIA GeForce RTX 2080 Ti, each with 11 GB of memory. In the case of the hybrid approach we used Python 3.7, PyTorch-1.8.1, scikit-learn-0.23.1, and an NVIDIA TESLA V100 with 16 GB memory. Finally, for the end-to-end approach we used Tensorflow 1.4, Keras 2.1.5, sklearn 0.22.2, matplotlib 3.0.3 and an NVIDIA GTX 1080 Ti with 8 GB memory.

5. Discussion

This study originated during the first wave of infection in Italy occurring in early spring, 2020, when thousands of people arrived every day in hospitals. Despite their apparently similar conditions, some lived the infection as a seasonal flu while others rapidly deteriorated, making intensive care necessary. This situation is common worldwide and, to fight the pandemic, in the last months the whole scientific community has carried out relevant research efforts in different fields of knowledge.

Artificial intelligence is one of the scientific disciplines that has been attracting more attention, offering the possibility to process and extract knowledge and insights from the massive amount of data generated during the pandemic, and it has mostly impacted prediction, diagnosis and treatment. Within this context, large efforts have been directed towards the analysis of radiological images and, according to the analysis presented by Greenspan et al. (2020), detection of COVID-19 pneumonia in both CT and CXR (Minaee et al., 2020; Zhang et al., 2020) is the field where large research has been directed to. Recently, there has been growing interest in the development of AI models to predict the severity of the COVID-19 infections because of the pressure on the hospitals, where even during the second pandemic wave we have assisted to an increasing demand for beds in both ordinary wards and intensive care units. The few papers available in this field use CT images, but several guidelines and statements do not encourage the use of CT over CXR (Rubin et al., 2020) and, for several practical reasons, CXR imaging is used due to the difficulty of moving bedridden patients, the lack of CT machine slots, the risk of cross-infection, etc.

To deal with this issue, here we have investigated different AI approaches mining CXR examinations and clinical data to predict the prognosis of 820 patients, whose data come from a multicentre retrospective study including 6 Italian hospitals. The results provide to researchers and practitioners a baseline performance reference to foster further studies.

With reference to the results attained by the AI approaches that process the clinical data only, using a normalized unitary scale Fig. 6 shows the rate each clinical descriptor was included in the selected feature subset by the RFECV wrapper, distinguishing also per classifier used. The figure shows the cumulative results observed running both the 10-fold and LOCO cross validation experiments. We opted for this cumulative representation since the trend is very similar in both the experiments. Furthermore, the readers can find in the figure also the set of biomarkers providing the best performance shown in the first section of Tables 4 and 5, which are denoted by reporting before an "*" or a "+" for 10-fold and LOCO cross validation experiments, respectively. Interestingly, Fig. 6 shows that age, LDH and O_2, were chosen in every fold for all the classifiers. If we used only such three descriptors, the average classification accuracy attained by learners in 10-fold and LOCO cross validation is equal to 0.74 ± 0.05 and to 0.70 ± 0.10, respectively. Moreover, sex, dyspnoea and WBC were always selected by the wrapper with the SVM and RF, whereas the D-dimer was always selected by the logistic regressor and by SVM. Oppositely, heart failure and cough were scarcely selected. Notably, some features such as LDH, D-dimer and SaO_2 were selected very frequently despite a high fraction of data was obtained by imputation (see Table 2). We deem that is mostly related to the strong differences in the distributions of these features between the two classes.

Fig. 6 also shows in dark blue the clinical feature relevance estimated by the deep learning approach (DL series). In this case the feature relevance was estimated as the maximum across neurons of the absolute value of the weights in the perceptron first layer. Results have been averaged over cross-validation folds and repetitions and rescaled to the [0,1] interval in order to match the other three series. Comparing the results with those obtained with the RFECV wrapper, it is clear that the only feature with the maximum relevance for all approaches is LDH, while sex, dyspnoea, WBC and CRP present a score higher than 0.5 in all series. The impact of the other clinical attributes appears to vary significantly depending on the adopted approach. For example, a high value of D-dimer and WBC have shown to be an important risk factor for negative outcome (Henry et al., 2020; Petrilli et al., 2020; Zhang et al., 2020). Furthermore, D-dimer, WBC and other clinical features like dyspnoea and LDH are indicators of pulmonary compromise, infection, tissue damage (Li et al., 2020) and a prothrombotic state (Naymagon et al., 2020) respectively. Finally, from our first statistical analysis (Section 2.1) of the dataset and from the result is shown in Fig. 6 the patient gender showed to have an in important role in classifying the patient severity. The reasons behind this difference appear to be related to the stark difference in immune system responses, with females causing stronger immune responses to pathogens. This difference can be a major

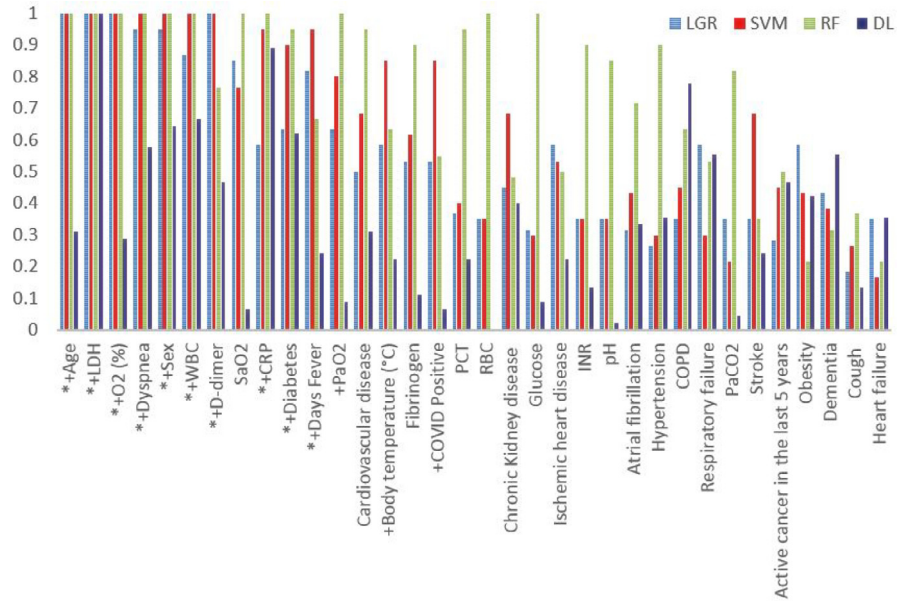


Fig. 6. Clinical feature importance represented by the rate each descriptor was selected by the RFECV wrapper during both the 10-fold and LOCO cross validation experiments using the three classifiers (LGR, SVM and RF series). The DL series represents feature importance estimated as the maximum absolute value of weights in the first layer of the perceptron of the DL network, after averaging over folds and repetitions and rescaling in the [0,1] interval. Moreover, the "*" or a "+" reported before each feature name means that it is included in the feature set used to get the best handcrafted results reported in the first section of Tables 4 and 5, respectively For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.

Table 4
Best recognition performance attained by each of the learning methods when the experiments were executed according to the 10-fold cross-validation (20 repetitions). In the second column, ML and DL stands for Machine-Learning and Deep Learning, respectively. The last column reports the learners providing the results shown here.

Input data	Approach	Accuracy	Sensitivity	Specificity	Learner
Clinical data	ML	0.757 ± 0.008	0.760 ± 0.007	0.754 ± 0.011	SVM
	DL	0.684 ± 0.019	0.753 ± 0.020	0.654 ± 0.012	MLP
CXR images	Handcrafted	0.658 ± 0.015	0.676 ± 0.016	0.638 ± 0.019	LGR
	Hybrid	0.728 ± 0.038	0.769 ± 0.072	0.680 ± 0.076	VGG-11 + RF
	End-to-end	0.742 ± 0.010	0.748 ± 0.019	0.738 ± 0.013	Resnet50
Clinical data and CXR images	Handcrafted	0.755 ± 0.007	0.758 ± 0.008	0.753 ± 0.013	SVM
	Hybrid	0.769 ± 0.054	0.788 ± 0.064	0.747 ± 0.059	GoogleNet + SVM
	End-to-end	0.748 ± 0.008	0.745 ± 0.017	0.751 ± 0.015	Resnet50 + MLP

Table 5
Best recognition performance attained by each of the learning methods when the experiments were executed according to the LOCO cross-validation. In the second column, ML and DL stands for Machine-Learning and Deep Learning, respectively. The last column reports the learners providing the results shown here.

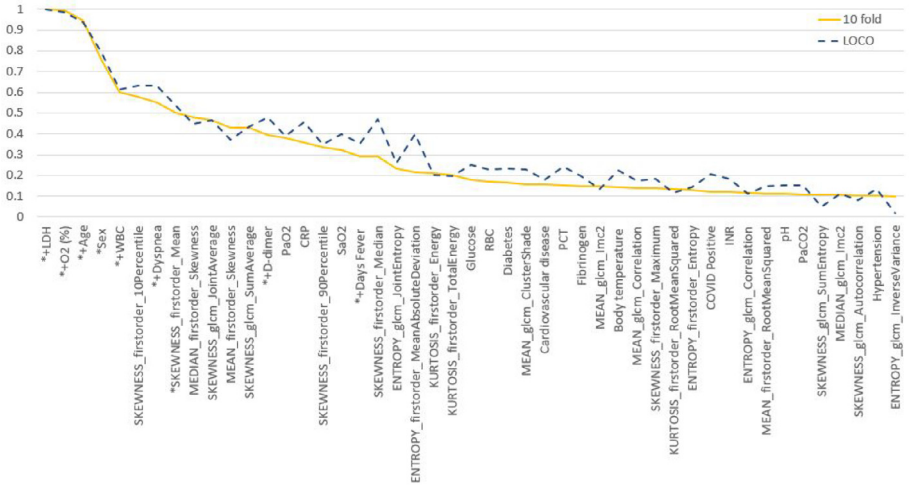
Input data	Approach	Accuracy	Sensitivity	Specificity	Learner
Clinical data	ML	0.734 ± 0.044	0.699 ± 0.158	0.795 ± 0.136	SVM
	DL	0.663 ± 0.016	0.709 ± 0.032	0.644 ± 0.018	MLP
CXR images	Handcrafted	0.625 ± 0.083	0.641 ± 0.159	0.644 ± 0.200	SVM
	Hybrid	0.693 ± 0.053	0.806 ± 0.161	0.549 ± 0.213	Vgg11 + SVM
	End-to-end	0.705 ± 0.010	0.720 ± 0.011	0.696 ± 0.015	Resnet50
Clinical data and CXR images	Handcrafted	0.752 ± 0.067	0.711 ± 0.165	0.824 ± 0.154	LGR
	Hybrid	0.743 ± 0.061	0.769 ± 0.189	0.685 ± 0.155	GoogleNet + LGR
	End-to-end	0.709 ± 0.005	0.734 ± 0.018	0.696 ± 0.009	Resnet50 + MLP

contributing factor to viral load, disease severity, and mortality. Furthermore, differences in sex hormone environments could also be a determinant of viral infections as oestrogen has immune-stimulating effects while testosterone has immune-suppressive effects (Pradhan and Olsson, 2020).

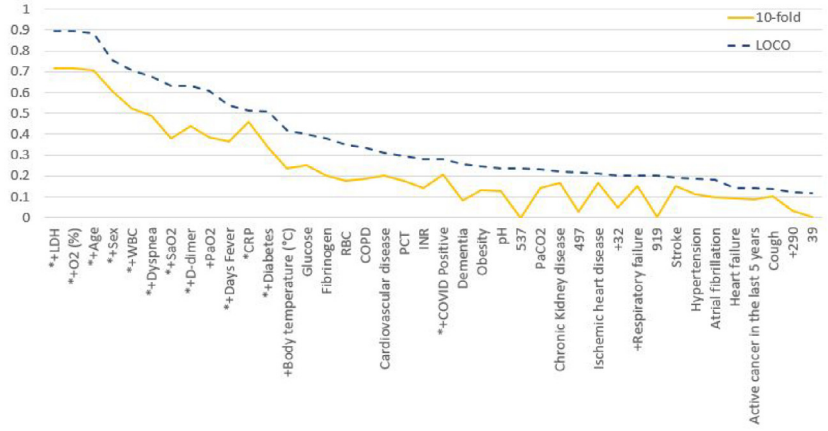
To deepen the use of semantic data as model input we take also into consideration the CXR radiological severity score proposed by Wong et al. (2020) and further investigated by Orsi et al. (2020). To this goal, an expert radiologist with more than 10 years of experience assigned such lung damage burden score to a cohort of 240 images randomly selected from the dataset. Then, this score is added to the clinical feature set as an additional image-derived feature. An SVM with the REFCV feature selection, which is the best performing architecture on clinical data as shown in Table 4, is used to classify the samples in 10-fold cross-validation with 20 repetitions according to the following three different experiments. First, to have a performance baseline, we ran again the experiment on this subset of images using the clinical features only (first row of Table 6); second, we test what happens using such score only (second row of Table 6); third, we ran another experiment using a feature set given by the clinical descriptors plus the radiological score (last row of the same ta-

Table 6
Recognition performance attained on a cohort of 240 images from the whole dataset when the human-based CXR radiological score proposed by Wong et al. (2020) and Orsi et al. (2020) is added to the clinical data.

Input data	Accuracy	Sensitivity	Specificity
Clinical data	0.728 ± 0.018	0.701 ± 0.039	0.758 ± 0.032
Only radiological score	0.718 ± 0.021	0.682 ± 0.032	0.750 ± 0.023
Clinical data + radiological score	0.719 ± 0.021	0.720 ± 0.030	0.724 ± 0.018



(a) Handcrafted approach



(b) Hybrid approach

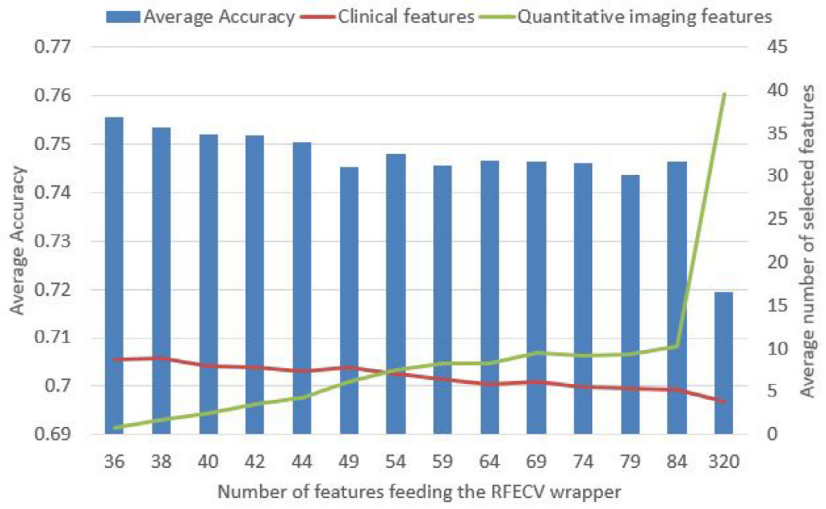
Fig. 7. Importance of clinical and handcrafted (panel A) or automatically learnt features (panel B) measured as the rate each descriptor was selected by the RFECV wrapper during the 10-fold and LOCO cross-validation experiments considering all the three classifiers employed. The *y* axis scale is normalized to one. Moreover, we add a "*" or a "+" before each feature name if it is included in the feature set used to get the best handcrafted or hybrid results reported in the last section of Tables 4 and 5, respectively.

ble). It is worth noting that in this last experiment the severity score is included in the set of selected features in 184 out of the 200 runs. The results show that the CXR severity score provides lower performance than the use of clinical descriptors only, regardless if it is used alone or in conjunction with such descriptors. Furthermore, the accuracies are not statistically different according to the Kruskal–Wallis test ($p = 0.545$). A Dunn's test with Bonferroni correction confirmed the result. This suggests that the use of a human-based score assessing lung damage burden is not beneficial.

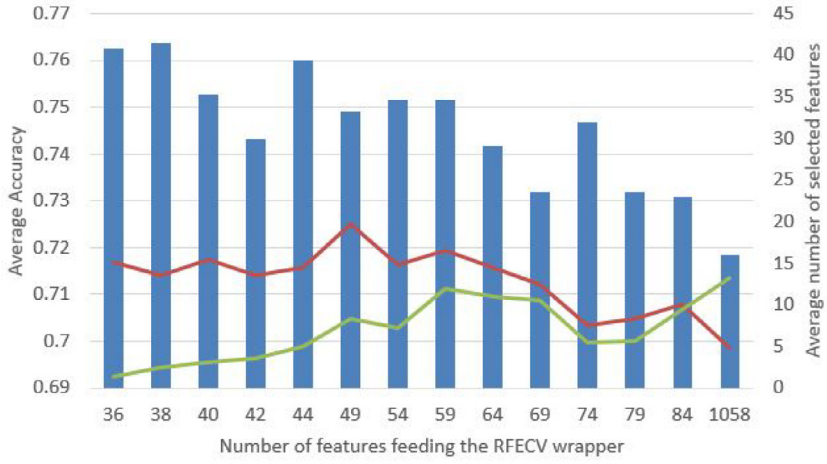
With reference to the results attained by the handcrafted approach, we found that the best results in terms of accuracy are statistically lower than those attained by the clinical descriptors ($p < 0.001$ and $p < 0.05$ for 10-fold and LOCO cross validation, respectively).

The approach that computes handcrafted features from the images also unfavourably compares with those using CNNs. Indeed, comparing with the hybrid and the end-to-end DL approaches we found that the performances are statistically different in both the 10-fold and LOCO cross validation tests, as we always got $p < 0.05$. No statistically different performances were found, instead, between the end-to-end and hybrid approaches.

Furthermore, Fig. 7a shows the feature importance of the 40 most selected handcrafted descriptors by the RFECV wrapper during the experiments in 10-fold and LOCO cross-validation. The feature relevance is computed as the number of times a feature is included in the selected subset during all the experiments performed using all the learners and, for the sake of clarity, all the values are normalized in [0,1]. The plot shows that the top-five descriptors most frequently detected as discriminative are clini-



(a) Handcrafted approach



(b) Hybrid approach

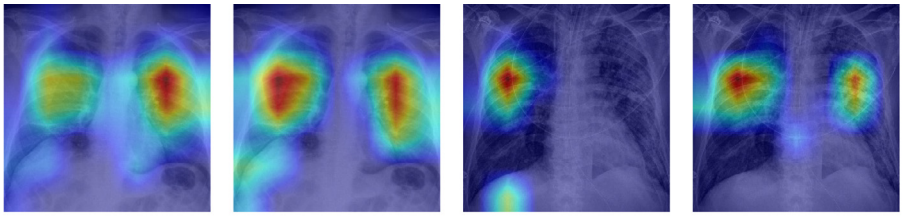
Fig. 8. Variation of the average classification accuracy (blue bars) with the number of features feeding the RFECV wrapper. The red and green curves show the number of clinical and texture features selected by the RFECV wrapper, respectively. The experiments plotted here refer to the best results shown in Table 4 integrating clinical and imaging features for the handcrafted (panel A) and hybrid approach (panel B). For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.

cal measures, followed by several texture measures almost equally distributed between the first- and second-order measures. For the sake of completeness, in this figure on the x-axis we add a "*" or a "+" before each feature name when it is included in the feature set used to get the best results by combining handcrafted measures from CXR images and clinical data, reported in the last section of Tables 4 and 5, respectively.

We now analyse how the performance of the handcrafted approach vary with the number of features selected by the coarse step, which fed the fine selection based on the RFECV method, as described in Section 3.6. To this end, Fig. 8a reports on the x-axis the number of features in input to the RFECV, which ranges from 36 (i.e. 34 clinical plus 2 texture measures) up to 84 (i.e. 34 clinical plus 50 texture measures), plus the last value where the RFECV received all the clinical and all the image features.[5] The bars show the average classification accuracy (y-axis, left side), while the curves in red and green show the average number of clinical and handcrafted texture features selected by the RFECV, respectively (y-axis, right side). As already noticed in Table 4, the use of texture measures does not improve the performance attained using the clinical descriptors; this is also confirmed by observing that, as the number of input features increases, the wrapper tends to select more imaging biomarkers than clinical ones, dropping the performance. This may remark the importance of using both clinical and imaging biomarkers since they may provide complementary information: while the former, and especially comorbidities, refers to the functional reserve of the patient, the latter may quantify the actual impact on the lungs. Indeed, fit patients with severe infection and damage are as likely as unfit-patients with less severe infections to have a poor prognosis. Although not reported, similar considerations can be derived in the case of LOCO cross-validation where we noticed that the best performances are attained by an almost balanced number of clinical and imaging features.

With reference to the results attained by the hybrid approach on the CXR images only, we found that the best results are statistically lower than those attained by the clinical descriptors for 10-fold cross validation ($p < 0.001$) but no differences were found with LOCO cross validation ($p = 0.24$). Among the three learners used with the hybrid approach, the best results with 10-fold cross validation are obtained with RF ($p < 0.001$, Kruskal–Wallis and

[5] The experiments plotted in Fig. 8a refer to the best results shown in Table 4 integrating clinical and imaging features by the handcrafted approach.



(a) Mild class, all neurons (b) Mild class, 40 most selected neurons (c) Severe class, all neurons (d) Severe class, 40 most selected neurons

Fig. 9. Two examples of the activation maps provided by the Grad-CAM approach, using all the neurons in the dense layer of the CNN dense layer or all the 40 neurons selected by the RFECV wrapper.

Dunn's test) while no differences were found with LOCO validation. Furthermore, comparing with a full DL approach, the hybrid provide lower performance ($p < 0.05$ and $p = 0.24$ for 10-fold and LOCO cross validation, respectively), suggesting that a fully connected layer better exploits the automatic features computed by the convolutional layers of the CNNs. As in Fig. 7a, we show in Fig. 7b the feature importance of the 40 most selected descriptors by the RFECV wrapper during the experiments in 10-fold and LOCO cross-validation using the GoogleNet. The plot shows that the features most frequently detected as discriminative are clinical measures with some neurons of the dense layer that, although few in number, permits to improve the classification accuracy. To deepen the results, Fig. 9 shows how much the selected neurons contribute to the network predictions. To this goal, we first depict the regions of input that are important for outputs provided by the CNN (panels a and c in the figure) by applying the Gradient-weighted Class Activation Mapping (Grad-CAM) approach Selvaraju et al. (2017). In a nutshell, Grad-CAM uses the class-specific gradient information flowing into the final convolutional layer to produce a coarse localization map of the relevant regions in the image. Next, we ran the same algorithm using only the 40 neurons in the dense layer that were mostly selected by the RFECV wrapper (panels b and d in the same figure). The visually inspection of the figure shows that the regions activated by the 40 neurons cover most of the areas activated by the whole dense layer, confirming that the wrapper correctly identifies the neurons carrying most of the information. Finally, as in Fig. 8a and b shows that using all the features automatically learnt does not help the learner improving the accuracy, whilst a limited and small number of descriptors is beneficial.

Still with reference to the handcrafted approach, in Section 3.3 we reported that lung segmentation performance attained by the U-Net are satisfactory to recover the bounding boxes of the lungs, which are then given as input to the CNN performing the feature computation. To deepen this issue we investigate if the use of lung regions automatically segmented differently impacts the final performance with respect to the use of lung masks manually delineated. To this goal, we use the best model combination of the hybrid approach shown in Tables 4 and 5 where, however, the CNN is applied on lung regions manually segmented. The results are reported in Supplementary Table 7, and they show that the performance attained are almost the same or even slightly worse than those achieved using automatically segmented lung regions (Tables 4 and 5). Furthermore, we also find that such differences are not statistically significant according to Wilcoxon's test with $p = 0.05$, except for one case where the results correspond to a classifier with lower performance than those reported before, and obtained using the automatic segmentation.

The end-to-end deep learning approach was built with the intuition that, through joined training of clinical information and images, it would be possible to generate better features for classification than by using either source alone. This idea was at least partly vindicated, as the classification results for the fully-DL approach proved higher for the combined approach than from either single source in the 10-fold cross-validation scenario ($p = 0.02$) and, arguably, for the LOCO case, as well ($p = 0.06$).

Furthermore, as already mentioned, classification accuracy from images alone is better than other methods, confirming the well-established finding that CNNs are powerful approaches for image classification. Oppositely, a neural network-based approach suffered particularly in achieving good performance with clinical information as inputs. The most likely cause for this underperformance is the fact that the clinical information structure is not standard and, therefore, it was impossible to adopt already tested network models and, more importantly, to pre-train the network on other datasets. It is likely that further fine-tuning of the design and training procedure of the custom multi-layer perceptron adopted for clinical-info classification could further improve results both with this specific input source, as well as for the combined model. A similar result is expected with an increase in size of the available dataset, as this section of the network did not undergo any pre-training, as mentioned above.

We now delve into the effect of pre-training the CNNs on another CXR dataset, that would help the models learning modality-specific feature representations, as already mentioned in Sections 3.7 and 3.8. In the case of the hybrid approach we did not find any significant difference in the performance achieved with and without this step: for instance, when using pre-training and the same configuration providing the best result using CXR images in 10-fold cross-validation (fourth line in Table 4) we get an accuracy equal to 0.712 ± 0.047. Similarly, in the case of using clinical data and CXR images (seventh line in Table 4) we get an accuracy equal to 0.768 ± 0.036 when we pre-trained the CNN. In the case of the end-to end approach, as already mentioned in Section 3.8, pre-training led to more consistent, but not better results. In fact, accuracy over 20 repetitions averages to 0.742 ± 0.001 when pre-training on Mooney's database, while training directly from the ImageNet weights result in 0.741 ± 0.002 average accuracy; median values are not statistically different (Wilcoxon's rank sum test p-value: 0.45), but results are more consistent across repetitions when pre-training is introduced (standard deviation of accuracy across repetitions is almost halved).

Let us now discuss how performances vary when Anterior Posterior (AP) and Posterior-Anterior (PA) projections are used. To this goal, we measure the performance for each of the best learners reported in Tables 4 and 5 distinguishing between the accuracies achieved on AP and PA images belonging to the different cross-validation instances of the test sets or to the different centres involved. The results, detailed in the supplementary material, show that in most of the cases the accuracies obtained on the AP images are larger than those achieved on the PA images; nevertheless, the AP scores are not so larger than the average results shown in Tables 4 and 5. Although AP images were mainly used for acquisitions of bedridden patients using portable machines producing a poorer quality image when compared with a

PA chest radiograph performed in a dedicated radiography facility (Cleverley et al., 2020), we deem that the larger accuracies they provide are due to their larger prior probability than the PA projections in the available repository (Table 1). This should support the proposed approach because the use of AI has revealed the possibility to predict the prognosis of the patients even in spite of the limitations of AP CXR scans, e.g. their more difficult interpretation and the sub-optimal imaging resulting from patient's positioning that may reduce inspiratory effort (Cleverley et al., 2020).

Finally, let us now focus on the population characteristics, where we found interesting reports on the age and gender distribution. Women were both less and older, suggesting that they become less ill and suffer from more serious conditions at an older age than men; also the women mortality was lower, as 72% were male confirming the male mortality reported in China (73%) by Chen et al. (2020). The male-related susceptibility and the higher male-mortality rate was also reported by Borges do Nascimento et al. (2020), who analyses the data of 59254 patients from 11 different countries. The second main finding was that 87% of patients had at least one comorbidity (Fig. 3), suggesting that, in most cases, the conditions leading to hospitalisation occur in patients with coexisting disorders. The most common disease (in 45% of cases) was hypertension, confirming the results reported by Yang et al. (2020a), who meta-analysed the data of 1576 infected patients from seven studies and reported an hypertension prevalence of 21%.

This study has also some limitations. First, patient enrolment was not globally randomized but instead conducted to populate the two classes with a roughly homogeneous number of cases. This implies that training data do not reflect the true a-priori probabilities of the target classes. On the other hand, sampling within the mild and severe classes is unbiased because patients were randomly enrolled.

Although this may bias the estimate of classification accuracy, there exist methods for adjusting the outputs of a trained classifier with respect to different prior probabilities without having to retrain the model, even when these probabilities are not known in advance (Latinne et al., 2001).

A further limitation but, from another point of view a key feature of the study, is the lack of full standardization of images and clinical data in the dataset. The dataset was built during spring 2020 when Italy was under lockdown and Italian hospitals and doctors were under pressure due to the huge amount of patients requiring hospitalization. Under these circumstances, full standardization of clinical data collection and images acquisition could not be achieved, and we decided to collect CXR images gathered under any conditions and all the clinical data most commonly acquired at the time of patients hospitalization. This led to a dataset that reflects these circumstances with many missing values among clinical data and images acquired with unstandardized clinical protocol (i.e. patient position and breath holding) and various devices. Although on the one side this may represent a limitation, on the other side it may be an advantage because this dataset could challenge the AI community on real data collected under critical circumstances. Another limitation may be the ever-changing landscape of the pandemic. Compared to the first wave, in many countries, and especially in Europe, the second wave has been characterised by younger patients with early symptoms in the emergency department. This may suggest to periodically re-train the learners to follows the disease evolution, or to investigate the use of methods able to cope with concept drifts (Lu et al., 2018).

5.1. Take-home messages and future works

In this preliminary analysis the use of image-derived data provide reduced predictive performance improvement with respect to the use of clinical data alone. The analysis of clinical data, instead, showed that a number of measures have robust predictive potential. Clinical data such as Age, LDH, O_2, Dyspena, Sex, WBC, D-dimer, SaO_2 are consistently selected across the different validation conditions and classifiers tested in this work, representing a set of biomarkers that can have impact in the clinical practice helping physicians and care-managers planning the bed allocations. Furthermore, the use of experts based score of lung damage burden was not found to be beneficial.

The poor standardization of images in the dataset could be a possible cause of the results attained here, as it has led to a classification problem hard to be addressed by the tested approaches. Indeed, beyond the variability introduced by non-standardized acquisition conditions such as patient positions and imaging device, the number of various medical devices, metal objects and other artefacts (e.g. pacemakers, catheters, prosthesis, etc.) that can be observed within the field of view are additional sources of difficulty for the learners. This suggests to further explore the dataset by using methods that can manage such variability, for instance by disregarding those images not meeting some quality criterion that can be learnt in parallel with the classification task. With reference to the approaches investigated here, deepening how data augmentation impact network training, performing ablation studies on the hybrid approach as well as on network sizes for the end-to-end DL procedure are future directions of investigation. Furthermore, to improve the quality of DNNs we deem that joint learning could be another direction of investigation, enabling the possibility to extract correlated information across clinical and imaging data to the used to enforce the network weights to be shared across these networks.

In conclusion, the dataset presented here is unique, offering a large number of CXR for prognostic purposes, placing side by side with similar efforts that use CT images (Chassagnon et al., 2020), and making available even more images. While this repository lets the machine learning community to challenge their methods with poorly standardized data, the efforts to collect a large repository cannot be afforded by such community, asking for the collaboration of researchers from different backgrounds, clinicians, and institutes. This is what we have started to carry out by promoting this long-term initiative that is still collecting other images to be added to the repository used here, as described in the next section. Furthermore, the quantitative results reported offer a preliminary evaluation of the prognostic performance attainable using AI approaches spanning from the use of handcrafted image descriptors to a fully automatic approach based on DNNs. The use of AI in this domain can open the chance to develop fast and low-cost clinical protocols, and the future availability in the repository of more annotated images will foster further research to obtain consistent results from the imaging contribution to the outcome prediction.

Data availability

The dataset generated and analysed in this study is publicly available to members of the scientific community upon request at *aiforcovid.radiomica.it*.[6] Beyond that, we encourage other hospitals and clinical centres to join the network to share their data; in this case, contacts for data sharing are also available on the website. As mentioned in Section 2, the dataset contains the CXR images, the clinical data listed in Table 2, the labels, the blind association between each image and the acquisition centre, and the acquisition information. The manual segmentation masks mentioned in Section 3.6 are not publicly available at the time this manuscript is submitted, and they will be added later on.

[6] We offer Reviewers the possibility to navigate within the dataset using this anonymous link.

Declaration of Competing Interest

Authors declare that they have no conflict of interest.

CRediT authorship contribution statement

Paolo Soda: Conceptualization, Methodology, Resources, Formal analysis, Writing – original draft, Writing – review & editing, Visualization, Project administration, Supervision. **Natascha Claudia D'Amico:** Conceptualization, Methodology, Software, Validation, Formal analysis, Resources, Data curation, Writing – original draft, Writing – review & editing. **Jacopo Tessadori:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Supervision. **Giovanni Valbusa:** Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Valerio Guarrasi:** Methodology, Software, Investigation, Data curation, Writing – original draft. **Chandra Bortolotto:** Conceptualization, Investigation, Writing – original draft, Writing – review & editing, Resources, Data curation, Visualization. **Muhammad Usman Akbar:** Methodology, Software, Validation, Investigation, Data curation, Writing – review & editing, Visualization. **Rosa Sicilia:** Methodology, Writing – original draft, Writing – review & editing, Formal analysis, Visualization. **Ermanno Cordelli:** Methodology, Writing – original draft, Writing – review & editing, Formal analysis, Visualization. **Deborah Fazzini:** Data curation. **Michaela Cellina:** Investigation, Writing – original draft, Writing – review & editing, Resources. **Giancarlo Oliva:** Investigation, Supervision. **Giovanni Callea:** Investigation, Resources, Data curation. **Silvia Panella:** Investigation, Resources, Data curation. **Maurizio Cariati:** Investigation, Resources, Supervision. **Diletta Cozzi:** Investigation, Resources, Data curation. **Vittorio Miele:** Investigation, Resources, Supervision. **Elvira Stellato:** Investigation, Resources, Data curation. **Gianpaolo Carrafiello:** Investigation, Resources, Supervision. **Giulia Castorani:** Investigation, Resources, Data curation. **Annalisa Simeone:** Investigation, Resources, Supervision. **Lorenzo Preda:** Conceptualization, Methodology, Writing – original draft, Supervision. **Giulio Iannello:** Writing – review & editing, Resources, Supervision. **Alessio Del Bue:** Conceptualization, Writing – review & editing, Supervision, Project administration. **Fabio Tedoldi:** Writing – review & editing, Supervision. **Marco Alí:** Conceptualization, Methodology, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Diego Sona:** Conceptualization, Methodology, Validation, Writing – original draft, Writing – review & editing, Supervision, Project administration. **Sergio Papa:** Writing – review & editing, Supervision.

Acknowledgements

The authors wish to thank Amazon Web Services (AWS) and the AWS Diagnostic Development Initiative for the support in putting in place the data management infrastructure. We also acknowledge FSTechnology SpA, which offered GPU usage for the experiments to Universita' Campus Bio-Medico.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2021.102216.

References

American College of Radiology, 2020. ACR Recommendations for the use of Chest Radiography and Computed Tomography (CT) for Suspected COVID-19 Infection. https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Recommendations-for-Chest-Radiography-and-CT-for-Suspected-COVID19-Infection. Online; accessed November, 30 2020.

Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., Tao, Q., Sun, Z., Xia, L., 2020. Correlation of ChestCT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. Radiology 296 (2), 200642.

Arcuri, A., Fraser, G., 2013. Parameter tuning or default values? An empirical investigation in search-based software engineering. Empir. Softw. Eng. 18 (3), 594–623.

Chassagnon, G., Vakalopoulou, M., Battistella, E., Christodoulidis, S., Hoang-Thi, T.-N., Dangeard, S., Deutsch, E., Andre, F., Guillo, E., Halm, N., et al., 2020. AI-driven quantification, staging and outcome prediction of COVID-19 pneumonia. Med. Image Anal. 67, 101860.

Chen, T., Wu, D., Chen, H., Yan, W., Yang, D., Chen, G., Ma, K., Xu, D., Yu, H., Wang, H., et al., 2020. Clinical characteristics of 113 deceased patients with coronavirus disease 2019: retrospective study. BMJ 368, 1–12.

Gozes, O., Frid-Adar, M., Greenspan, H., Browning, P. D., Zhang, H., Ji, W., Bernheim, A., Siegel, E., 2020. Rapid AI development cycle for the coronavirus (COVID-19) pandemic: initial results for automated detection & patient monitoring using deep learning CT image analysis. arXiv preprint arXiv:2003.05037.

Cleverley, J., Piper, J., Jones, M.M., 2020. The role of chest radiography in confirming COVID-19 pneumonia. BMJ 370, 1–9.

Greenspan, H., Estépar, R.S.J., Niessen, W.J., Siegel, E., Nielsen, M., 2020. Position paper on COVID-19 imaging and AI: from the clinical needs and technological challenges to initial AI solutions at the lab and national level towards a new era for AI in healthcare. Med. Image Anal. 66, 101800.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Mach. Learn. 46 (1–3), 389–422.

Haralick, R.M., Shanmugam, K., Dinstein, I.H., 1973. Textural features for image classification. IEEE Trans. Syst. Man Cybern. 3 (6), 610–621.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Henry, B.M., De Oliveira, M.H.S., Benoit, S., Plebani, M., Lippi, G., 2020. Hematologic, biochemical and immune biomarker abnormalities associated with severe illness and mortality in coronavirus disease 2019 (COVID-19): a meta-analysis. Clin. Chem. Lab. Med. (CCLM) 58 (7), 1021–1028.

Hesamian, M.H., Jia, W., He, X., Kennedy, P., 2019. Deep learning techniques for medical image segmentation: achievements and challenges. J. Digit. Imaging 32 (4), 582–596.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., Keutzer, K., 2016. SqueezeNet: alexnet-level accuracy with 50x fewer parameters and < 0.5 MB model size. arXiv preprint arXiv:1602.07360.

Imlab-UIIP. Lung Segmentation (2D). https://github.com/imlab-uiip/lung-segmentation-2dR. Online; accessed 19 October 2020.

Jaeger, S., Candemir, S., Antani, S., Wáng, Y.-X.J., Lu, P.-X., Thoma, G., 2014. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quant. Imaging Med. Surg. 4 (6), 475.

Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al., 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172 (5), 1122–1131.

Krizhevsky, A., 2014. One weird trick for parallelizing convolutional neural networks. arXiv preprint arXiv:1404.5997.

Latinne, P., Saerens, M., Decaestecker, C., 2001. Adjusting the outputs of a classifier to new a priori probabilities may significantly improve classification accuracy: evidence from a multi-class problem in remote sensing. In: ICML, 1, pp. 298–305.

Leeuwenberg, A.M., Schuit, E., 2020. Prediction models for COVID-19 clinical decision making. Lancet Digit. Health 2 (10), e496–e497.

Li, C., Ye, J., Chen, Q., Hu, W., Wang, L., Fan, Y., Lu, Z., Chen, J., Chen, Z., Chen, S., et al., 2020. Elevated lactate dehydrogenase (LDH) level as an independent risk factor for the severity and mortality of COVID-19. Aging (Albany NY) 12 (15), 15670.

Liu, C., Cao, Y., Alcantara, M., Liu, B., Brunette, M., Peinado, J., Curioso, W., 2017. TX-CNN: detecting tuberculosis in chest X-ray images using convolutional neural network. In: 2017 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 2314–2318.

Lu, H., Stratton, C.W., Tang, Y.-W., 2020. Outbreak of pneumonia of unknown etiology in Wuhan, China: the mystery and the miracle. J. Med. Virol. 92 (4), 401–402.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., Zhang, G., 2018. Learning under concept drift: a review. IEEE Trans. Knowl. Data Eng. 31 (12), 2346–2363.

Ma, N., Zhang, X., Zheng, H.-T., Sun, J., 2018. ShuffleNet v2: practical guidelines for efficient CNN architecture design. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 116–131.

Minaee, S., Kafieh, R., Sonka, M., Yazdani, S., Soufi, G.J., 2020. Deep-COVID: predicting COVID-19 from chest X-ray images using deep transfer learning. Med. Image Anal. 65, 1–9.

Mockus, J., 2012. Bayesian Approach to Global Optimization: Theory and Applications, 37. Springer Science & Business Media.

Mooney, P., 2017, Chest X-ray images (Pneumonia). https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia. Online; accessed 16 October 2020.

Moons, K.G., Wolff, R.F., Riley, R.D., Whiting, P.F., Westwood, M., Collins, G.S., Reitsma, J.B., Kleijnen, J., Mallett, S., 2019. Probast: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann. Intern. Med. 170 (1), W1–W33.

Borges do Nascimento, I.J., Cacic, N., Abdulazeem, H.M., von Groote, T.C., Jayarajah, U., Weerasekara, I., Esfahani, M.A., Civile, V.T., Marusic, A., Jeroncic, A., et al., 2020. Novel coronavirus infection (COVID-19) in humans: a scoping review and meta-analysis. J. Clin. Med. 9 (4), 941.

Naymagon, L., Zubizarreta, N., Feld, J., van Gerwen, M., Alsen, M., Thibaud, S., Kessler, A., Venugopal, S., Makki, I., Qin, Q., et al., 2020. Admission D-dimer levels, D-dimer trends, and outcomes in COVID-19. Thromb. Res. 196, 99–105.

Orsi, M.A., Oliva, G., et al., 2020. Feasibility, reproducibility, and clinical validity of a quantitative chest X-ray assessment for COVID-19. Am. J. Trop. Med. Hyg. 103 (2), 822–827.

Penny, W.D., Friston, K.J., Ashburner, J.T., Kiebel, S.J., Nichols, T.E., 2011. Statistical Parametric Mapping: The Analysis of Functional Brain Images. Elsevier.

Petrilli, C.M., Jones, S.A., Yang, J., Rajagopalan, H., O'Donnell, L., Chernyak, Y., Tobin, K.A., Cerfolio, R.J., Francois, F., Horwitz, L.I., 2020. Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in New York City: prospective cohort study. BMJ 369, 1–15.

Pradhan, A., Olsson, P.-E., 2020. Sex differences in severity and mortality from COVID-19: are males more vulnerable? Biol. Sex Differ. 11 (1), 1–11.

Radiological Society of North America, (2018) RSNA Pneumonia Detection Challenge. https://www.rsna.org/en/education/ai-resources-and-training/ai-image-challenge/RSNA-Pneumonia-Detection-Challenge-2018. Online; accessed 15 November 2020.

Rajaraman, S., Siegelman, J., Alderson, P.O., Folio, L.S., Folio, L.R., Antani, S.K., 2020. Iteratively pruned deep learning ensembles for COVID-19 detection in chest X-rays. IEEE Access 8, 115041–115050.

Rajaraman, S., Sornapudi, S., Alderson, P.O., Folio, L.R., Antani, S.K., 2020. Analyzing inter-reader variability affecting deep ensemble learning for COVID-19 detection in chest radiographs. PLoS One 15 (11), e0242301.

Ross, B., 2014. Mutual information between discrete and continuous data sets. PLoS One 9, e87357. doi:10.1371/journal.pone.0087357.

Rubin, G., Ryerson, C., Haramati, L., Sverzellati, N., Kanne, J., et al., 2020. The role of chest imaging in patient management during the COVID-19 pandemic: a multinational consensus statement from the fleischner society. Chest.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2018. MobileNetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520.

Schiaffino, S., Tritella, S., Cozzi, A., Carriero, S., Blandi, L., Ferraris, L., Sardanelli, F., 2020. Diagnostic performance of chest X-ray for COVID-19 pneumonia during the SARS-CoV-2 pandemic in Lombardy, Italy. J. Thorac. Imaging 35 (4), W105–W106.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626.

Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K.-i., Matsui, M., Fujita, H., Kodera, Y., Doi, K., 2000. Development of a digital image database for chestradiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. Am. J. Roentgenol. 174 (1), 71–74.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9.

Vancheri, S.G., Savietto, G., Ballati, F., Maggi, A., Canino, C., Bortolotto, C., Valentini, A., Dore, R., Stella, G.M., Corsico, A.G., et al., 2020. Radiographic findings in 240 patients with COVID-19 pneumonia: time-dependence after the onset of symptoms. Eur. Radiol. 1, 6161–6169.

Wong, H.Y.F., Lam, H.Y.S., et al., 2020. Frequency and distribution of chest radiographic findings in patients positive for COVID-19. Radiology 296 (2), E72–E78.

Wynants, L., Van Calster, B., Bonten, M.M., Collins, G.S., Debray, T.P., De Vos, M., Haller, M.C., Heinze, G., Moons, K.G., Riley, R.D., et al., 2020. Prediction models for diagnosis and prognosis of COVID-19 infection: systematic review and critical appraisal. Br. Med. J. 369, 1–16.

Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500.

Xu, X., Guo, Q., Guo, J., Yi, Z., 2018. DeepCXray: automatically diagnosing diseases on chest X-rays using deep neural networks. IEEE Access 6, 66972–66983.

Yan, F., Huang, X., Yao, Y., Lu, M., Li, M., 2019. Combining LSTM and densenet for automatic annotation and classification of chest X-ray images. IEEE Access 7, 74181–74189.

Yang, X., Yu, Y., Xu, J., Shu, H., Liu, H., Wu, Y., Zhang, L., Yu, Z., Fang, M., Yu, T., et al., 2020. Clinical course and outcomes of critically ill patients with SARS–CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. Lancet Respir. Med. 8 (5), 475–481.

Yang, J., Zheng, Y., Gou, X., Pu, K., Chen, Z., Guo, Q., Ji, R., Wang, H., Wang, Y., Zhou, Y., 2020. Prevalence of comorbidities and its effects in patients infected with SARS-CoV-2: a systematic review and meta-analysis. Int. J. Infect. Dis. 94, 91–95.

Yue, H., Yu, Q., Liu, C., Huang, Y., Jiang, Z., Shao, C., Zhang, H., Ma, B., Wang, Y., Xie, G., et al., 2020. Machine learning-based CT radiomics method for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: a multicenter study. Ann. Transl. Med. 8 (14), 1–7.

Zagoruyko, S., Komodakis, N., 2016. Wide residual networks. arXiv preprint arXiv:1605.07146.

Zanardo, M., Schiaffino, S., Sardanelli, F., 2020. Bringing radiology to patient's home using mobile equipment: a weapon to fight COVID-19 pandemic. Clin. Imaging 68, 99–101.

Zhang, K., Liu, X., Shen, J., Li, Z., Sang, Y., Wu, X., Zha, Y., Liang, W., Wang, C., Wang, K., et al., 2020. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. Cell 181 (6), 1423–1433 e11.

Zhang, L., Yan, X., Fan, Q., Liu, H., Liu, X., Liu, Z., Zhang, Z., 2020. D-dimer levels on admission to predict in-hospital mortality in patients with covid-19. J. Thromb. Haemost. 18 (6), 1324–1329.

Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2018. UNet++: a nested U-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 3–11.