End-to-end learning for semiquantitative rating of COVID-19 severity on Chest X-rays

Alberto Signoroni\textsuperscript{a,1,*}, Mattia Savardi\textsuperscript{a,1}, Sergio Benini\textsuperscript{a}, Nicola Adami\textsuperscript{a}, Riccardo Leonardi\textsuperscript{a}, Paolo Gibellini\textsuperscript{b}, Filippo Vaccher\textsuperscript{b}, Marco Ravanelli\textsuperscript{b}, Andrea Borghesi\textsuperscript{b}, Roberto Maroldi\textsuperscript{b}, Davide Farina\textsuperscript{b}

\textsuperscript{a}Department of Information Engineering – University of Brescia, Brescia, Italy
\textsuperscript{b}Department of Medical and Surgical Specialties, Radiological Sciences, and Public Health – University of Brescia, Brescia, Italy

Abstract

In this work we designed an end-to-end deep learning architecture for predicting, on Chest X-rays images (CRX), a multi-regional score conveying the degree of lung compromise in COVID-19 patients. Such semiquantitative scoring system, namely Brixia-score, was applied in serial monitoring of such patients, showing significant prognostic value, in one of the hospitals that experienced one of the highest pandemic peaks in Italy. To solve such a challenging visual task, we adopt a weakly supervised learning strategy structured to handle different tasks (segmentation, spatial alignment, and score estimation) trained with a “from part to whole” procedure involving different datasets. In particular, we exploited a clinical dataset of almost 5,000 CXR annotated images collected in the same hospital. Our BS-Net demonstrated self-attentive behavior and a high degree of accuracy in all processing stages. Through inter-rater agreement tests and a gold standard comparison, we were able to show that our solution outperforms single human annotators in rating accuracy and consistency, thus supporting the possibility of using this tool in contexts of computer-assisted monitoring. Highly resolved (super-pixel level) explainability maps were also generated, with an original technique, to visually help the understanding of the network activity on the lung areas. We eventually tested the performance robustness of our model on a variegated public COVID-19 dataset, for which we also provide Brixia-score annotations, observing good direct generalization and fine-tuning capabilities that favorably highlight the portability of BS-Net in other clinical settings.

Keywords: COVID-19 severity assessment
Chest X-Rays
Semiquantitative rating
End-to-end learning
Convolutional Neural Networks

1. Introduction

Among the most critical aspects of the ongoing COVID-19 pandemic, there is the saturation of healthcare facilities, caused by the high contagiousness of Sars-Cov-2, and by the significant rate of complications of the disease. Nearly 20% of patients affected by COVID-19 develop severe respiratory symptoms that may require hospitalization (WHO [2020]). Diagnostic imaging, specifically chest X-ray (CXR) and computed tomography (CT) can play an essential role in the management of these patients (Rubin et al. [2020]). CXR is a fast and accessible diagnostic modality, which may be easily brought to the
patient’s bed, either in the emergency department and in the hospital wards, as acknowledged by the Italian Society of Medical Radiology [SIRM 2020]. Although CXR is known to be hampered by low sensitivity, significantly inferior to CT, the extensive use of CT can be difficult to sustain for organizational reasons and raises concerns for the radiation dose delivered to a population that includes a large number of young subjects. In addition, the lower sensitivity of CXR may be compensated by the high prevalence of viral pneumonia, particularly at the peak of the epidemic [Manna et al. 2020; Rubin et al. 2020].

In Italy, starting February 2020, the region of Lombardy was the epicenter of the outbreak of COVID-19 and Brescia was among the most severely involved provinces. At the peak of the epidemic, the local university hospital (ASST Spedali Civili di Brescia) admitted up to 900 COVID-19 patients. In such an overburdened scenario, CXR was chosen as the workhorse. It must be emphasized that the clinical utility of any imaging test depends only partially on the intrinsic qualities of the diagnostic modality. An essential part actually relates to the ability of the radiologist to convey precisely and in the most immediate way the information provided by the exam. In the highly critical conditions that ASST Spedali Civili di Brescia had to face, radiologists sought to improve the communication with the referring clinicians adopting a semiquantitative approach to CXR reading. With this aim, a semiquantitative scoring system, namely Brixia-score, was designed and implemented in the routine [Borghesi and Maroldi 2020]. Brixia-score classifies lung involvement in viral pneumonia providing a net score integrated with elements that describe the type of lung lesion and its spatial localization (see Figure 1a and Sec 3.1). Since in the routine turnaround of the shifts in a radiology department CXR exams are inevitably reported by different radiologists, the Brixia-score system was crucial for monitoring the course of lung disease on serial exams. Codification of the site and type of lung lesions, in fact, made the comparison of CXR exams faster and significantly more consistent. In the recent literature, CXR scores based on the subdivision of lungs in regions [Borghesi and Maroldi 2020; Toussie et al. 2020], were applied in serial monitoring of COVID-19 patients and showed significant prognostic value [Borghesi et al. 2020a;b].

Digital technologies [Ting et al. 2020], artificial intelligence (AI), and data science [Latif et al. 2020] are concurrently giving an incredible contribution to this unprecedented pandemic management. AI-based solutions have the potential to expand the role of chest imaging beyond diagnosis, to facilitate risk stratification, disease progression monitoring, and trial of novel therapeutic targets, in the global race to contain and treat COVID-19 [Kundu et al. 2020]. For example, the presence of AI-assisted first preliminary readings, or second opinions, can be determinant in the speed-up and efficiency of decision-making procedures, especially in critical conditions like the ones characterizing the management of health facilities overload. In such a context, Deep Learning (DL) based computer-aided radiological image interpretation systems [Shi et al. 2020] can be an important asset for more effective handling of the above described clinical tasks. However, the design of data-driven methods is critical in terms of reliability and safety of use even for sophisticated automated tools [Filice and Ratwani 2020; Reyes et al. 2020], it requires accurate data collection and preparation [Willemink et al. 2020] and, to overcome the proof of concept stage, it requires order of thousands of images of adequate quality [Manna et al. 2020]. The early availability of public datasets of CXR images from COVID-19 subjects – although relatively small – catalyzed the research on the assisted COVID-19 diagnosis (e.g., differential diagnosis with other bacterial/viral forms of pneumonia). However, very few works applying AI to CXR in the context of COVID-19 diagnosis considered large enough datasets and rigorous experimental methods [Castiglioni et al. 2020; Maguolo and Nanni 2020; Tartaglione et al. 2020]. Moreover, almost no work has been done in other directions, despite the fact that severity assessment on CXR has been highlighted as one of the highest reasonable research efforts to be pursued in the field of AI-driven COVID-19 radiology [Laghi 2020].

In this work, we focus on AI-driven assessment of the severity of lung involvement on CXR images in the context of COVID-19 pandemic. In particular, we propose an end-to-end representation learning solution for the evaluation of the degree of severity of COVID-19 pneumonia based on a multi-purpose deep neural network and characterized by the estimation of the semiquantitative multi-valued Brixia-score [Borghesi and Maroldi 2020], that expresses the degree of lung compromise. Assessing the severity of the pathology, especially for CXR,
is a highly difficult visual task with high feature variability (from subtle findings to heavy lung impairment), which requires a large dataset. We had the opportunity to collect and analyze a huge amount of CXR (about 5,000 radiograms) from confirmed COVID-19 patients, corresponding to the entire amount of images taken during both triage and patient monitoring in sub-intensive and intensive care units of our hospital during one month, between March 4th and April 4th 2020. This constitutes an adequate amount of data to train a network able to tackle the challenging task of providing an automatic semiquantitative assessment of the COVID-19 severity. To this end, we designed an original DL-network architecture able to learn and execute different tasks such as data normalization, lung segmentation, geometric alignment, feature extraction, and multi-valued score estimation. This solution evidences good self-attentive behavior, and bias-free learning, with performance adequate to the challenge of the visual task and the need to operate in a weakly supervised scenario, coherently to radiologist performance in executing the same task. Moreover, this allowed us to conduct a portability test on public datasets, hypothesizing that this solution could be used and improved in other centers, or for other related goals.

The rest of the paper is structured as follows: in Section 2 a brief review is proposed in relation to the current efforts on AI and CXR especially the ones recently issued in this pandemic period, trying to recognize potentials, issues, and open challenges; in Section 3 we justify the focus of our work on AI-assisted CXR monitoring over a large dataset, recall the characteristics of the Brixia-score and present the whole image analysis architecture; Section 4 describes the collected dataset and the other public datasets we utilized, as well as the additional score annotations we produced for public COVID-19 CXRs; in Section 5 design and details of the proposed architecture and its training are provided; in Section 6 evaluation and discussion of the obtained performance is given; the section also addresses the degree of portability/robustness of the approach. The whole architecture source code is released for research purposes, and we also contribute to public datasets with two rounds of Brixia-score annotations.

2. Related Work

A recent review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19 can be found in Shi et al. (2020), where authors structured previous work according to different tasks, e.g., contactless imaging workflow, image segmentation, disease detection, radiomic feature extraction, etc. The use of convolutional neural networks (CNN) to analyze Chest X-ray for presumptive early diagnosis and better patient handling based on signs of pneumonia, has been proposed by Oh et al. (2020) at the very beginning of the outbreak, when a systematic collection of a large CXR dataset for deep neural network training was still problematic. As the epidemic was spreading, the increasing availability of COVID-19 CXR datasets has polarized almost all the research efforts on AI-based interpretation of these images for diagnosis-oriented studies. In the period from late February to mid-May 2020 we found more than 60 works focusing in this direction (most of them in a pre-print format). As it is impossible to mention them all, the reader can refer to some early reviewing efforts, as the one in Shi et al. (2020). Many of these approaches aggregate multiple datasets and consider multiple classes: normal, COVID-19, and other types of pneumonia (Linda Wang and Wong, 2020; Oh et al., 2020; Li et al., 2020). Most methods exploit available public COVID-19 CXR datasets (Kalkreuth and Kaufmann, 2020). Constantly updated collections of different public COVID-19 CXR datasets are curated by the authors of (Linda Wang and Wong, 2020; Cohen et al., 2020). Prior to COVID-19, large CXR datasets have been released and also used in the context of open challenges for the analytics of different types of pneumonia Wang et al. (2017) or pulmonary and cardiac (cardiomegaly) diseases Irvin et al. (2019). These datasets are usually exploited as well in works related to COVID-19 to complete the case categories, especially when the focus is on differential diagnosis among other types of viral and bacterial pneumonia.

2https://github.com/lindawangg/COVID-Chexpert-master/docs/COVIDx.md
3https://github.com/ieee8023/covid-chestxray-dataset
4https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data
5https://stanfordmlgroup.github.io/competitions/chexpert/
Class balancing is a relevant aspect in all these methods (Pereira et al., 2020), as well as the comparison of AI performance with those of independent radiologists (Murphy et al., 2020). Many of these data-driven works are interesting in principle and can be inspirational, but almost all would require and benefit from extensions and validation on much larger datasets. This is indeed the case for most of the studies issued in this emergency period, often based on sub-optimal experimental designs, since the numerous still unknowns factors about COVID-19 severely undermine the external validity and generalizability of the performance of diagnostic tests (Sardanelli and Di Leo, 2020). Cautions about a radiologic diagnosis of COVID-19 infection driven by deep learning have been also expressed by Laghi (2020), who states that a more interesting application of AI in COVID-19 infection is to provide for a more objective quantification of the disease, in order to allow the monitoring of the prognostic factors (i.e., lung compromising severity) for appropriate and timely patient treatment.

Beyond the fact that CXR modality should be more appropriate for monitoring and severity assessment than for primary detection, other issues severely affect the previous studies which employed CXR for COVID-19 diagnosis purposes (Burlacu et al., 2020). Working with datasets of a few hundred images, when analyzed with deep architectures, often results in severe overfitting and can encounter issues generated by unbalanced classes when larger datasets are used to represent other classes (e.g. other kinds of pneumonia). In (Maguolo and Nanni, 2020) it is pointed out that many CXR based systems seem to learn to recognize more the characteristics of the dataset rather than those related to COVID-19. This effect can be overcome, or at least mitigated, by working on more homogeneous and larger datasets, as in (Castiglioni et al., 2020), or by preprocessing the data as in (Pereira et al., 2020), or by including lung segmentation before analysis (Tartaglione et al., 2020).

Since we also perform lung segmentation, it is important to mention that this task has been recently witnessing a convergence towards encoder-decoder architectures, such as in the U-Net framework (Ronneberger et al., 2015), and in the fully convolutional approach found in Long et al. (2015). All these approaches share skip connections as a common key element, originally found in ResNet (He et al., 2016) and DenseNet architectures (Huang et al., 2017). The idea behind skip connections is to combine coarse-to-fine activation maps from the encoder within the decoding flow. In doing so, these models are capable of efficiently using the extracted information from different abstraction layers. In Zhou et al. (2018), a nested U-Net is presented, bringing the idea of skip connection to its extreme. This approach is well suited to capture fine-grained details from the encoder network, and exploding them in the decoding branch.

Last, in critical sectors like healthcare, the lack of understanding of complex machine-learned models is hugely problematic. Therefore, explainable AI approaches able to reveal to physicians where is directed the model attention (Karim et al., 2020; Oh et al., 2020; Rajaraman et al., 2020; Reyes et al., 2020), are always desirable.

Despite some evidenced issues, there is still an abundant ongoing effort on AI-driven approaches for COVID-19 detection based on CXR analysis. Conversely, we did not find AI-driven solutions for disease monitoring and lung severity assessment based on CXR, although this modality, for the abovementioned reasons, is part of the routine practice in many institutions, like the one from which this study originates.

3. Background, aims and contribution

A quantitative assessment of lung involvement in COVID-19 pneumonia by means of DL methods was proposed on CT in (Huang et al., 2020; Gozes et al., 2020), and can be highly relevant in both diagnostic and prognostic contexts. However, routine use of CT for patient monitoring is unsustainable for reasons related to radioprotection, prevention of infection spread, and difficulty of moving the most compromised patients.

Hence, a strong indication for the use of CXR emerges. Since CXR is less sensitive than CT, and since working on 2D projections is not recommended for quantitative assessment of affected volumes, it is not advisable to develop a purely quantitative measure (e.g., a 2D area measure) to approach AI-driven assessment of COVID-19 severity based on CXR. A semiquantitative scoring system can instead leverage the sensitivity of CXR as well as the ability of radiologists to detect COVID-19 pneumonia and communicate in an effective way its severity according to an agreed severity scale.
3.1. Brixia-score on CXR

The multi-region 6-valued Brixia-score was designed and implemented in routine reporting by the Radiology Unit 2 of ASST Spedali Civili di Brescia (Borghesi and Maroldi, 2020), and later validated for risk stratification on a large population in (Borghesi et al., 2020b). According to it, lungs in anteroposterior (AP) or posteroanterior (PA) views, are subdivided into six zones, three for each lung, as shown in Figure 1(a):

- Upper zones (A and D): above the inferior wall of the aortic arch;
- Middle zones (B and E): below the inferior wall of the aortic arch and above the inferior wall of the right inferior pulmonary vein (i.e., the hilar structures);
- Lower zones (C and F): below the inferior wall of the right inferior pulmonary vein (i.e., the lung bases).

Whenever it is difficult to identify some anatomical landmarks, due to technical reasons (for example bedside CXR in critical patients), it is acceptable to divide each lung into three equal zones. For each zone, a score (ranging from 0 to 3) is assigned, based on the detected lung abnormalities:

- 0: no lung abnormalities;
- 1: interstitial infiltrates;
- 2: interstitial (dominant), and alveolar infiltrates;
- 3: interstitial, and alveolar (dominant) infiltrates.

The six scores may be then aggregated to obtain a Global Score in the range [0, 18]. During the peak period, the Brixia-score has been systematically used to report CXR in COVID-19 patients. Examples of scores assigned to different cases are showcased in Figure 1(b-d).

3.2. Semiquantitative score estimation with an AI system

An automatic assessment of a semiquantitative prognostic score may seem at first sight easier than other tasks, such as differential diagnosis or purely quantitative severity evaluations. However, it turns out to be quite the opposite. Quantitative tasks are defined on a definite target, e.g., opacity volumetric quantification in CT, as an indicator of disease progression (Gozes et al., 2020), with the possibility to have ground truth information (for training) fully coherent to the quantitative task. On the other side, image interpretation for diagnosis can be seen as a qualitative task (dichotomous presence/absence judgment) where again ground truth information can be collected (from clinical records) and is fully coherent with the task. In the case of Brixia-score, the annotation is on a quantitative scale but related to a mostly qualitative task, which is characterized by a certain degree of subjectivity (hence the definition of semiquantitative). In this case, at least two major critical aspects arise. First, the difficulty of establishing a ground truth information, since subjective differences in scoring were expressed by 15 radiologists that contributed to annotations, characterized by the visual assessment and qualification of the presence of, sometimes subtle, abnormalities. Second, the localization of these findings remains implicit and related to the visual attention of the specialist in following...
anatomical landmarks (without any explicit localization information indicated, nor lung segmentation provided). This results in the difficulty to define reference spatial information usable as ground truth and produces a lack of completeness of the annotations with respect to the task. Moreover, the visual task related to severity assessment is inherently challenging, since CXR findings in COVID-19 may be extremely variable (from no or subtle signs to extensive changes that modify the anatomical patterns and borders), and the quality of the information conveyed by the images may be impaired due to the presence of medical devices, as well as to sub-optimal patient positioning. These factors, if not handled, can impact in an unpredictable way on the reliability of an AI-based interpretation. This concomitant presence of quantitative and qualitative aspects, on a visually difficult task, makes the 6-valued Brixia-score estimation on CXR particularly challenging.

3.3. Weakly-supervised self-attentive framework

Intra- and inter-operator subjectivity in scoring assignments, and the lack of detailed localization information, determines the need to operate in a weakly supervised framework (Zhou, 2017) and led to a the design of an end-to-end DL solution. This would facilitate a multi-valued and semiquantitative estimation of the Brixia-score capable to operate in such a tough scenario. Far from constituting a weakness, weakly supervised approaches have demonstrated to be highly valuable methods to leverage available knowledge in medical domains (Xu et al., 2014; Wang et al., 2017; Karimi et al., 2019; Bontempi et al., 2020; Tajbakhsh et al., 2020). In addition, we favor a system able to express self-attentive behaviors, since the score annotations do not include explicit spatial annotations. The network must therefore be able to select the correct region of the lungs for the 6 partial score predictions. A functional scheme of the proposed solution is shown in Figure 2. The end-to-end network architecture, represented as a group of four hexagons (and described in detail in Section 5.1), comprises a shared backbone, a lung segmentation part, a spatial registration part, and a classification part (for the 6-valued score estimation) and interacts for training, validation, and testing phases with 4 CXR datasets (described in Section 4):

- the collected dataset at ASST Spedali Civili of Brescia, including 4707 clinical CXR from COVID-19

Figure 2. Overview of the proposed method.
patients;
• a public dataset of about 200 COVID-19 CXR from several centers;
• a public CXR dataset annotated for lung segmentation;
• and a synthetic dataset for image alignment (rotation, scale, translation).

4. Dataset

The training and testing of the multi-task DL architecture of Figure 2 requires the use of multiple datasets, all described in this section.

4.1. Brixia COVID-19 dataset

We collected a large dataset of CXR images that is a faithful picture of what really happened during the pandemic peak in ASST Spedali Civili of Brescia, and contains all the variability originating from a real clinical scenario. It includes 4707 CXR images of COVID-19 subjects, acquired with both Computed Radiography (CR) and Digital X-ray (DX) modalities, in AP or PA projection, and retrieved from the institutional RIS-PACS system. All image reports include the Brixia-score as a string of six digits indicating the scores assigned to each region, A to C for the right lung, and D to F for the left one. The Global Score is simply the sum of the 6 numbers. Scores were assigned by a team of about 15 radiologists operating in different radiology units of the hospital, with a very wide range of years of experience and different specific expertise in imaging of the chest. All images were collected and anonymized, and their usage for this study had the approval of the local Institutional Review Board (number #300170016). We subsequently requested an authorization to release the whole anonymized dataset for research purposes (currently pending). The main characteristics of the dataset are summarised in Table 1.

4.2. Public COVID-19 dataset

As an additional dataset, we exploit the public repository by Cohen et al. (2020), which contains CXR images of patients which are positive or suspected of COVID-19. In order to contribute to such public dataset, two expert radiologists, a board-certified staff member and a trainee with 22 and 2 years of experience respectively, produced the related Brixia-score annotations for CXR in this collection, exploiting labelbox, an online solution for labelling. After discarding problematic cases (e.g., images with a significant portion missing, too small resolution, the impossibility of scoring for external reasons, etc.), the final dataset is composed of 194 CXR, completely annotated according to the Brixia-score system.

Beyond increasing the scope and value of the public CXR dataset, by providing additional annotations for tasks different than diagnosis, we will make use of the annotated COVID-19 repository to demonstrate the robustness and the portability of the proposed solution, even on the complex case mix of CXR present in this collection (different provenance, variable image quality,...).

4.3. Segmentation datasets

We exploit different segmentation datasets in order to pre-train the Nested-Unet module of the proposed architecture. In particular, we used Montgomery County (Jaeger et al., 2014), Shenzhen Hospital (Stirenko et al., 2018), and JSRT database (Shiraishi et al., 2000). We

In order to contribute to such public dataset, we took into account the scope and value of the public CXR dataset, by providing additional annotations for tasks different than diagnosis, we will make use of the annotated COVID-19 repository to demonstrate the robustness and the portability of the proposed solution, even on the complex case mix of CXR present in this collection (different provenance, variable image quality,...).

Table 1
Brescia dataset details.

| Parameter       | Value                           |
|-----------------|---------------------------------|
| Modality        | CR (62%) - DX (38%)             |
| View position   | AP (87%) - PA (13%)             |
| Manufacturers   | Carestream, Siemens             |
| Image size      | (1056 ÷ 3050) x (1186 ÷ 3376)   |
| No. of images   | 4707                            |
| Training set    | 3313 images                     |
| Validation set  | 945 images                      |
| Test set        | 449 images                      |

---

6We downloaded a copy on May 11th, 2020.

7https://labelbox.com/
used the original training/test set splitting when present (as the case of the JSRT database), otherwise we took the first 50 images as test set, and the remaining as training set (see Table 2).

Table 2  
Segmentation datasets.

|                | Training-set | Test-set | Split  |
|----------------|--------------|----------|--------|
| Montgomery County | 88           | 50       | first 50   |
| Shenzhen Hospital | 516          | 50       | first 50   |
| JSRT database     | 124          | 123      | original  |
| Total             | 728          | 223      |          |

4.4. Alignment dataset

CXR acquired in clinical settings may have not the perfect level of magnification or alignment of the lungs. To avoid the inclusion of anatomical parts not belonging to the lungs in the AI pipeline, which would increase the task complexity or introduce unwanted biases, we integrated into the pipeline an alignment block. This exploits a synthetic dataset (used for on-line augmentation) composed of artificially transformed images from the segmentation dataset (see Table 3), including random rotations, shifts, and zooms, which is used in the pre-training phase.

Table 3  
Alignment dataset: synthetic transformations. The parameters refer to the implementation in Albumentation [Buslaev et al., 2020]. In the last column is expressed the probability of the specific transformation being applied.

| Parameters (up to) | Probability |
|--------------------|-------------|
| Rotation 25 degree | 0.8         |
| Scale 10%          | 0.8         |
| Shift 10%          | 0.8         |
| Elastic transformation alpha=60, sigma=12 | 0.2        |
| Grid distortion steps=5, limit=0.3 | 0.2        |
| Optical distortion distort=0.2, shift=0.05 | 0.2        |

5. Method

5.1. Proposed Model

Given the complexity of the addressed problem, we propose a novel multi-purpose architecture able, in an end-to-end fashion, to segment, align, and predict the Brixia-score of a given CXR. Network weights have been trained through an articulated procedure, involving the various considered datasets, that foresees a pre-training on the specific tasks, followed by a global fine-tuning.

A global scheme of the network structure, with the multi-resolution analysis stages, is depicted in Figure 4. The input image is initially processed by a cascade of convolutional blocks (referred to as the Backbone), which generates the pyramid of features used by the segmentation network to isolate the image areas corresponding to the lungs. The segmentation probability map in output is then used to estimate the alignment transformation which is required to normalize both the segmentation mask and the feature pyramid. The aligned features are finally used to estimate the Brixia-score matrix. A hard self-attention mechanism can also be applied by masking the aligned features with the corresponding segmentation mask (taken as soft-max probability map). This step has the advantage of switching off possible misleading regions outside the lungs, favoring the flow of relevant regions only. Hereby, we call BS-Net-HA the network with the hard self-attention scheme, and BS-Net-SA the soft version that does not exploit the segmentation mask as an additional attention mechanism.

The Backbone, compactly represented in Figure 4 (in yellow) by the convolutional blocks cascade, is the most critical component of the proposed model, because it is used both as the encoder section of the U-Net, and as the feature extractor for the Feature Pyramid Network of the classification branch. To identify the best solution, with respect to the goals of this work, we tested different backbones among the state-of-the-art, i.e., ResNet [He et al., 2016], VGG [Simonyan and Zisserman, 2014], DenseNet [Huang et al., 2017], and Inception [Szegedy et al., 2017].

Segmentation is performed by a nested version of Unet, also called Unet++ [Zhou et al., 2018], a specialized architecture oriented to medical image segmentation (in blue in Figure 4). It is composed of an encoder-decoder structure where the encoder branch, the Backbone, exploits a nested interconnection, to the decoder. These skip pathways are able to squeeze and extract the important information present in medical images, providing a smart reuse of already extracted features.

Alignment is realized through a spatial transformer net-
Figure 3. Exemplification of the alignment through the sampling grid produced by the transformation matrix, and its application either to the segmentation mask and the feature maps. On the right, the hard-attention mechanism and the ROI Pooling operation.

work (Jaderberg et al., 2015) able to estimate the alignment matrix, so that it learns to center, and correctly zoom, the lungs (see Figure 3). This alignment block, pre-trained on the synthetic alignment dataset, produces a self-attentive mechanism useful to propagate the right portion of the lungs, through the ROI (Region Of Interest) pooling, toward the final classification stage.

Scoring head exploits the ideas of Feature Pyramid Networks (Lin et al., 2017) (FPN) for the generation of multi-scale feature maps. As depicted in Figure 4 we combine feature maps that come from various levels of the network, therefore with different semantic information at various resolutions. The presence of the alignment block, in conjunction with the optional masking, produces input feature maps that are well focused on the specific area of interest. Eventually, the output of the FPN layer flows in a series of convolutional blocks to retrieve the output map (3 × 2 × 4, i.e., 3 rows, 2 columns and 4 score classes). The classification is done by a final Global Average Pooling layer and a SoftMax activation.

5.2. Training

The Loss function we use for training, is a sparse categorical cross entropy (SCCE) with a (differentiable) mean absolute error (dMAE) contribution:

\[ L = \alpha \cdot SCCE + (1 - \alpha) \cdot dMAE \]  

where \( \alpha \) controls how much weight is given to SCCE and dMAE, which are defined as follows:

\[ SCCE = -\frac{1}{C} \sum_{c} y_c \log(\hat{y}_c) \]  

\[ dMAE = \frac{1}{C} \left\| y - \sum_{c} e^{\beta y_c} \sum_{j} e^{\beta y_j} c \right\| \]  

where \( y \) is the golden standard (Brixia-score), \( \hat{y} \) is the predicted one, and \( c \in [0, 1, \ldots, 3] \) is the score class. To make the mean absolute error differentiable (dMAE), \( \beta \) is chosen to be an arbitrary large value.

The selection of such loss function is coherent with the choice to configure the Brixia-score problem as a joint multi-class classification and regression. Tackling our score estimation as a classification problem allows to associate to each score a confidence value: this can be useful to either produce a weighted average, or to introduce a quality parameter that the system, or the radiologist, can take into account.

The training procedure, due to the complexity of the proposed network, takes place at several stages, as depicted in Figure 4:

1. Training of the Unet++ using the segmentation dataset with its lung segmentation masks;
2. Training of the alignment block using the synthetic alignment dataset (semi-supervised setting);
3. Training of the classification portion (Scoring head) on Brescia dataset, while blocking the weights of the remainder; (Backbone/Segmentation/Alignment)
4. Complete fine-tuning on our Brixia-dataset, making all weights (about 20 Million) trainable.

The network hyper-parameters then undergo a selection that maximizes the MAE score on the validation set.
5.3. Preprocessing

One of the most important ways of mitigating the differences present in a dataset, thus potentially solving unwanted biases that could impinge the training procedure, is a pre-processing step able to normalize the appearance of the CXR. All data are directly imported from DICOM files, consisting in 12-bit gray-scale images, and mapped to float32 between 0 and 1. We then apply sequentially: an adaptive histogram equalization (CLAHE, clip: 0.01); a median filtering to cope with noise (kernel size: 3); a clipping outside the 2nd and 98th percentile.

5.4. Implementation details

The network architecture has an input size of 512×512. The selected backbone is a ResNet-18 [He et al., 2016], because it has the best trade-off between expressiveness of the extracted features and memory footprint (as in the case of the input size). In the network, we use the rectified linear unit (ReLU) activation functions for the convolutional layer of the backbone, and the Nested-U-Net, while the Swish activation function by Ramachandran et al. [2017] is used for the remaining blocks. We extensively make use of online augmentation throughout the learning phases. In particular, we apply all the geometric transformations described in Section 4.4 plus random brightness and contrast, as well. Moreover, we randomly flip images horizontally (and the score, accordingly) with a probability of 0.5. We exploit, for training purposes, two machines equipped with Titan® V GPUs. We train the model by jointly optimizing the sparse categorical cross-entropy function and MAE, with a selected \(\alpha = 0.7\). Convergence was achieved after roughly 6 hours of training (80 epochs), using Adam [Kingma and Ba, 2014] with an initial learning rate of \(3 \times 10^{-2}\), halving it on flattening. Furthermore, we set the batch size to 8.

Other technical details not included here, for the sake of readability, can be found either in Figure 4, or in the documentation accompanying the source code distribution.8

---
8The code and the trained models will be released after publication on the project page on GitHub: [http://github.org/BrixIA](http://github.org/BrixIA). Updated info, additional material and results related to this project can be already found on the same page. The authorization to share the whole annotated Brixia-dataset has been submitted to our Ethical Committee and is
5.5. Explainability maps

To evaluate whether the network is predicting the Brixia-score on the basis of correctly identified lung areas, we need a method capable of generating maps with a sufficiently high resolution. Unfortunately, with the chosen network architecture, the popular Grad-CAM \cite{Selvaraju2017} approach generates poorly localized, and spatially blurred, visual explanations of activation regions, as it happens for example in \cite{Oh2020}, where a patch-based solution (although far from our approach) resulted beneficially. For these reasons, we designed a novel method, loosely inspired by the initial phases of the LIME \cite{Ribeiro2016} approach. Starting from the segmented and aligned lungs, we first generate a partitioning scheme by Quick-shift \cite{Vedaldi2008} into $N$ super-pixels. Then, for each of the $N$ super-pixel, we create an image with the corresponding super-pixel set to zero value (super-pixel-masked image), and for such $N$ images we predict the 6-valued Brixia-score. We then accumulate the differences with respect to the predicted values considering the whole original images for each class on the specific super-pixel. Given $S$ the set of super-pixels, the output explanation map $E$ is obtained as:

$$E = \sum_{i}^{\lvert S \rvert} S_i \cdot (p_i - p_0) \quad (4)$$

where $p_0$ is the probability map obtained with the original input, while $p_i$ is the probability associated to the super-pixel-masked image.

Intuitively, the obtained map highlights the region that most accounts for the final outcome. Examples can be appreciated in Figure 2 (bottom-right) as well as in Figure 9: the more intense the color, the more important the region for the score decision. In particular, we can see that the severe score in the bottom region of the right lung in Figure 2 is mainly due to the bottom left part of it.

6. Results and discussion

In the following we illustrate a variety of results from the experimental phases of our work. We refer to the presented deep network architecture as Brixia-score Network (BS-Net), which is used in one of its three configurations:

1. Hard attention (BS-Net-HA): the soft-max segmentation mask (from 0 to 1) is used as a (product) weighting mask for the CXR image;
2. Soft attention (BS-Net-SA): the segmentation mask is only used for spatial normalization and alignment of extracted features, but not to mask the CXR image;
3. Ensemble decision (BS-Net-Ens): two realizations of the two previous configurations are used to make the final prediction by averaging their output probabilities. The two employed realizations are the best model with respect to the validation set, and the obtained model after the last training iteration.

Additionally, we include the results provided by ResNet-18 (the backbone of our BS-Net), which is one among the most adopted architectures in previous studies on CXRs in the context of COVID-19.

6.1. Learning phase

Figure 5 shows the training curves (training set, and validation set parameters) tracking Intersection of Union (IoU) performance, also known as Jaccard index, for the segmentation and the alignment stage of the BS-Net-HA, and in terms of Mean Absolute Error (MAE) on the set of lung sectors for the classification stage (Brixia-score estimation). Convergence behaviors are clearly visible, with a consolidating residual distance between training and validation curves in the case of the alignment and classification tasks. This will be justified later.

6.2. Lung segmentation

Table 4 reports the results in terms of Dice coefficient and IoU of our implementation on the test set of the segmentation dataset, after both dedicated Unet++ and fine-tuning global trainings.

The performance of our segmentation stage, trained on large public datasets (see Section 4.3), substantially confirms and shows totally comparable results, with respect to both hybrid \cite{Candemir2014} and DL based \cite{Zhou2018,Frid-Adar2018,Oh2020} state-of-the-art methods \cite{Candemir2019}.

\[ \text{currently pending.} \]
6.3. Alignment

The estimation of a spatial normalization transform (translation, rotation, zoom) is used to correctly identify the 6 lung regions of interest on both the image and segmentation mask. After training on the synthetic dataset described in Section 4.4 we report fair realignment results: Dice coefficient = 0.873, IoU = 0.778. These performance need to be interpreted considering that transforms simulated in the training set are normally higher than issues we found in real data. From a qualitative point of view, after visual checking taken on all the 4707 images in the Brixia-dataset, no impairing anomaly (lung outside the normalized region) was observed after the combination of segmentation and alignment. Residual errors are typically in the form of slight rotation and zoom deviations which are not critical because well tolerable in terms of overall self-attentive behavior.

6.4. Brixia-score prediction

To evaluate the global performance of the proposed framework, we analyze the discrepancy between the network predictions and the scores assigned by the radiologist who originally annotated the CXR during the clinical practice. Such discrepancy is evaluated in terms of Mean Error (MEr), Mean Absolute Error (MAE), its standard deviation (SD), and the Correlation Coefficient (CC). The four networks considered for comparison are the three configurations of BS-Net, and ResNet-18.

Table 5 lists all performance values referred to each of the six regions of the Brixia-score (range [0-3]), to the average of single cells, and to the global score (range [0-18]). In a consistency assessment perspective, Figures 6 (top) and (bottom) show the confusion matrices for the four networks related to the score value assignments for single lung regions, and for their sum (Global Score), respectively. From Table 5 it clearly emerges fact that the ensemble decision strategy succeeds in combining the strengths of the soft and hard attention policies.

In principle, being the scoring scale defined on integers, a MAE measured on single regions below 0.5 could be interpreted as acceptable. This is a simplified reasoning, but interviewed radiologists, with hundreds of cases of experience of such semiquantitative rating system, also indicated, from a clinical perspective, ±0.5 as an acceptable error on each region of the Brixia-score, and ±2 as an acceptable error on the Global Score from 0 to 18. These indications are also based on the prognostic value and associated use of the score as a severity indicator that comes from the experimental evidence and clinical observations during the first period of its application (Borghesi et al., 2020a).

This is a first key of interpretation which casts a positive light on the obtained results and help to appreciate the differences between the network performance. In particular, for the ensemble decision network BS-Net-Ens we obtain an average MAE on cells of 0.44.

The fact that the ensemble decision strategy combines the strengths of the soft and hard attention policies deserves some further elaboration. In fact, if on one hand, removing the context is fine to avoid possible bias on the decision, on the other hand, the context info can help

| Backbone   | Dice coefficient | IoU  |
|------------|------------------|------|
| Unet++ ResNet-18 | 0.971 | 0.945 |
| Unet ResNet-18   | 0.969 | 0.941 |
when the segmentation is too restrictive. For example, retrocardiac consolidations that could be visible are removed from the lung segmentation and therefore their assessment is not allowed in the hard attention approach. This residual complementarity between the two options is exploited by the ensemble decision and explains the significant improvement.

What also emerges is that, despite largely adopted, a straightforward end-to-end approach by means of an all-learning ResNet-18 is always a worst option compared to the here proposed solutions. This clearly justifies the more structured approach we designed, which demonstrates to pay-off right from the soft-attention configuration. Another aspect that emerges looking at MER is that our architecture does not over/under-estimate, while ResNet-18 tends to overestimate.

A last qualitative evidence in favor of the use of a composite loss function, with an additional component related to the MAE (see Equation 1), is that the performance we obtain are unbiased and stable over several repetitions of the whole training process.

6.5. Inter-rater agreement

The above a-priori interpretation of ±0.5 as an acceptable error on a single region is clearly not sufficient. Therefore we need to thoughtfully assess the level of inter-rater agreement among human annotators to gain a reference measure of the inbuilt level of error in the training set. This is relevant since, being a source of error in our weakly supervised approach, such inter-rater agreement determines an implicit limit to the performance we can expect from the network.

Unfortunately, we cannot directly derive this information from the Brixia-dataset, since in origin we only had one Brixia-score annotation for each image. Therefore, we conducted specific tests to quantitatively assess this degree of inter-operator variability. To limit the overburden in this still overloaded period we asked to 4 radiologists to rate 1/3 of the test set of the Brixia-dataset (150 images). We name $R_0$, $R_1$, $R_2$, and $R_3$ the four radiologists that give additional scores on a set of 150 images from our test set. The four radiologists’ expertise is variegated to represent that of the whole staff experience: we have one resident at the 2nd year of training, and three staff radiologists with 9, 15 and 22 years of experience (reported numerical ordering of $R_i$ does not necessarily correspond to seniority order).
Table 5
Disparity measures between the 4 considered deep networks predictions and the Brixia-score in the COVID-19 patient reports (only blind test set results reported). Parameters are evaluated on each single lung portion (A-F), averaged on all the lung portions and on the global score (P-value \(< 0.00001\) everywhere).

|            | A   | B   | C   | D   | E   | F   | Avg. on regions | Global score |
|------------|-----|-----|-----|-----|-----|-----|-----------------|--------------|
| BS-Net-Ens | 0.169 | -0.038 | -0.056 | 0.125 | -0.045 | -0.192 | **-0.006** | **-0.036** |
| BS-Net-SA  | 0.107 | -0.087 | -0.171 | 0.082 | -0.129 | -0.343 | **-0.090** | **-0.541** |
| BS-Net-HA  | 0.156 | -0.147 | -0.085 | 0.107 | -0.016 | -0.238 | **-0.037** | **-0.223** |
| ResNet-18  | 0.356 | -0.038 | -0.056 | 0.125 | -0.045 | -0.192 | 0.100 | 0.601 |
| BS-Net-Ens | 0.459 | 0.448 | 0.412 | 0.374 | 0.459 | 0.494 | 0.441 | 1.728 |
| BS-Net-SA  | 0.499 | 0.501 | 0.506 | 0.408 | 0.499 | 0.566 | 0.496 | 1.846 |
| BS-Net-HA  | 0.481 | 0.477 | 0.481 | 0.370 | 0.488 | 0.532 | 0.471 | 1.826 |
| ResNet-18  | 0.543 | 0.486 | 0.506 | 0.452 | 0.584 | 0.530 | 0.517 | 1.951 |
| BS-Net-Ens | 0.604 | 0.524 | 0.540 | 0.541 | 0.574 | 0.609 | 0.565 | 1.429 |
| BS-Net-SA  | 0.638 | 0.560 | 0.579 | 0.576 | 0.594 | 0.634 | 0.597 | 1.514 |
| BS-Net-HA  | 0.613 | 0.575 | 0.583 | 0.552 | 0.594 | 0.616 | 0.589 | 1.505 |
| ResNet-18  | 0.657 | 0.579 | 0.591 | 0.632 | 0.657 | 0.601 | 0.619 | 1.710 |
| BS-Net-Ens | 0.675 | 0.779 | 0.731 | 0.682 | 0.737 | 0.672 | **0.713** | **0.862** |
| BS-Net-SA  | 0.635 | 0.733 | 0.675 | 0.633 | 0.718 | 0.636 | 0.672 | 0.847 |
| BS-Net-HA  | 0.665 | 0.742 | 0.679 | 0.662 | 0.722 | 0.645 | 0.686 | 0.845 |
| ResNet-18  | 0.598 | 0.739 | 0.667 | 0.562 | 0.655 | 0.643 | 0.644 | 0.842 |

Figure 7. Pairwise inter-rater results in terms of MAE (and SD). In the most right column (orange), the inter-rater results with predictions by BS-Net-Ens.

**Pairwise and averaged inter-rater variability.** In Figure 7, we assess the inter-rater agreement by listing MAE and SD values referred to all possible pairs of raters, including \(H\), and also BS-Net-Ens, as a further “virtual rater”. Looking at how the virtual rater BS-Net-Ens behaves (orange boxes in Figure 7) compared to how each human rater behaves with respect to each other colleague (blue boxes), we can conclude that our network is able to work equal or even better than each specific radiologist. For example, by considering \(H\) as a common reference, we have an equal performance only in the case of \(R_3\). At the same time, these values clearly evidence that there is a limit in what the network can learn, due to the assessed inter-rater variability. In Table 6 (top-center) we report the average performance on all rater pairs, and the performance of BS-Net-Ens recomputed on the reduced test set of 150 CRX. BS-Net-Ens (center) performs significantly better with respect to the global indicators coming from averaging all pairwise comparisons (top).

From a statistical point of view, both averaged two-raters Cohen’s kappa value and the multi-rater Fleiss’ kappa value, based on single cell scores is around 0.4. These values are indicative of a fair to moderate level of agreement. This validates the fact that the Brixia-score rating system (on four severity values) is a good trade-off between the two opposite needs of having a fine granularity of the rating scale, and a good inter-rater agreement. Moreover, this highlights once again the difficulty of this learning task.
Table 6

Results on the reduced test set (150 images) for inter-rater reasoning:
(top) averaged performance for all pairs in Figure 7; (center) performance of BS-Net-Ens recomputed on the reduced test set; (bottom) performance of BS-Net-Ens vs the Gold Standard.

|            | MEr | MAE   | SD   | CC   |
|------------|-----|-------|------|------|
| Average on all pairs of radiologists |     |       |      |      |
| Avg. on reg. | -0.131 | 0.528 | 0.614 | 0.736 |
| Global score | -0.784 | 2.592 | 1.965 | 0.835 |

| BS-Net-Ens vs H |     |       |      |      |
| Avg. on reg. | -0.019 | 0.452 | 0.575 | 0.754 |
| Global score | -0.113 | 1.847 | 1.553 | 0.834 |

| BS-Net-Ens vs Gold Standard |     |       |      |      |
| Avg. on reg. | -0.088 | **0.401** | 0.563 | 0.743 |
| Global score | 0.107 | 1.840 | 1.524 | 0.847 |

Gold Standard and performance (re)assessment. By exploiting the availability of multiple ratings (by $H, R_0, R_1, R_2$ and $R_3$) on the same dataset of 150 CRX, we produce a Gold Standard score based on a majority criterion. Then we recompute the BS-Net-Ens performance with respect to this more reliable reference. Table 6 (center-bottom) clearly shows a significant improvement with respect to the comparison versus $H$. This further demonstrates the quality and robustness of the Brixia-score estimation learning despite the inter-rater variability issues.

6.6. Ablation study

Taking as a reference the BS-Net-HA configuration we conducted an ablation study comprising two sets of experiments, which are carried out on the training (3313 CXRs) and on the validation (945 CXRs) sets.

Feature Extraction. A first question on which to reason is “which kind, and how complex, should be the feature map extraction leading to the Brixia-score estimation?” (see Scoring Head in Sec. 5.1). To seek an answer to this question, we compared our FPN-based solution with 1) a simplified, 2) a more complex, and 3) an increased spatial resolution configuration, and reported the results of this benchmark in Table 7. 1) Getting rid of a multi-scale approach, the score estimation can exploit the features extracted by the backbone head. While this produces an improved MAE on the training set, we observe a poorer generalization capability on the validation set. 2) We then tried a more articulated multi-scale approach based on the EfficientDet architecture [Tan et al., 2019], where BiFPN (Bidirectional FPN) blocks are introduced for an easy and fast multi-scale feature fusion; this however produces a worsening on both data sets. Therefore FPN confirms to be a good intermediate solution between 1) and 2). In 3) we added one resolution level to the FPN to allow the flow of $1024 \times 1024$ images. However, we clearly observed again a performance worsening (despite, for memory limitations, we did not succeeded in performing a complete fine-tuning). Therefore we considered the $512 \times 512$ resolution as well enough, and the pyramidal structure effective to fully benefit from information at various resolutions. Since we did not observe signs of overfitting related to an increased number of coefficients, one reason that can explain the last evidence on ideal spatial resolution is a slightly increased inter-rater variability for decisions involving subtle details, which might interfere with the possibility for the network to leverage an increased resolution.

Table 7

Performances on the training and validation set for different Feature Pyramid Networks (or lack of). (*) No complete fine-tuning due to memory limitation

|            | Train MAE (SD) | Val MAE (SD) |
|------------|---------------|--------------|
| FPN        | 0.455 (0.578) | **0.469** (0.583) |
| Backbone head | **0.395** (0.542) | 0.475 (0.592) |
| BiFPN      | 0.486 (0.593) | 0.504 (0.598) |
| FPN 1024 (*)| 0.498 (0.564) | 0.498 (0.589) |

Table 8

Performances on the training and validation set for different augmentation policies.

| Pre-proc. | Train MAE (SD) | Val MAE (SD) |
|-----------|---------------|--------------|
| No augmentation | y 0.306 (0.490) | 0.528 (0.610) |
| Bright. & contrast | y 0.341 (0.518) | 0.550 (0.628) |
| Geometric transf. | y 0.272 (0.460) | 0.541 (0.623) |
| All together | n 0.437 (0.585) | 0.571 (0.645) |
| All together | y 0.455 (0.578) | 0.469 (0.583) |
**Data Augmentation.** The second question for this ablation study is whether the pre-processing and augmentation policies we adopted are effective, or if there are some prevailing/redundant constituents. From Table 8 we clearly see that partial augmentation policies are not adequate, since they produce performance worsening on the validation set, and tend to create overfitting gaps between training and validation performances, while jointly they produce best results and no such a gap. Moreover, we can clearly appreciate the impact of the used equalization, which is able to correctly handle the multitude of differences present in the used dataset.

### 6.7. Errors and explainability visual analyses

The error distribution analysis, depicted in Figure 8, shows the prevalence of lower values errors on both single lung regions and Global Score estimations. The combined view gives another evidence that single-cell errors unlikely sum up as constructive interference.

In Figure 9 we illustrate three good predictions (top) and two among the worst cases (bottom). In particular, we report the lung segmentation and alignment, as well as the explainability maps obtained with the method described in Section 5.5. The latter has the advantage to clearly underline the regions that trigger a specific score with a colour. We decided to use a colour-map that goes from white in case of no impact, to the class colour (either orange, red, black for class 1,2,3, respectively) to imply high significance. This information can be used to increase trust in the outcome, and point in possible confounding situations. We had a valuable positive general feedback from radiologists that considered these maps as a possible complementary source of visual attention and of second level reasoning in a possible scenario of computer-assisted diagnosis. For example, in the first CXR in Figure 9 (top-left) the right areas of increased intensity, corresponding to the central and bottom left lung sectors (region E-F), offer a clear overview of the annotation/estimation agreement. The same holds true for the other full concordance cases of the top part of the figure. Instead, looking at bottom part of Figure 9 some more considerations are possible. The case on the bottom-left part, despite producing a correct Global Score (considered a positive fact by the radiologists, given the specific case), evidences both under- and over-estimations in single sectors. Reviewed by a senior radiologist this case evidenced some problematic areas even for the radiologists that is often called to express an average score about what it sees in different areas. For this reason the super-pixel explainability maps can be a useful tool of machine to human communication. In particular the sector A interested by the presence of an assisted ventilation device can determine a common problem for both machine an expert, but the machine proves to not diverge toward unduly high ranking, while the expert ruled in favor to the machine in the upper lung sector (D), while confirms a slight overestimation in sector E. In general, we can appreciate the fact that external equipment do not harm the prediction, nor produces unwanted biases as in Figure 9 (top-middle, bottom-left, bottom-right). The last case (bottom right) evidences a segmentation error on the right lung, since part of the colon (filled by air, pulling up the diaphragm and compressing the lung) enters in the segmentation mask, determining a wrong evaluation of the corresponding area. However, the radiologist positively considered the fact that the system sees a difference of two levels between the upper lung portions, which the expert considered a more coherent judgment at a second view.

### 6.8. Portability tests on public COVID-19 datasets

The aggregate public CXR dataset, described in Section 4.2, has been judged as inherently well representative with respect to the various manifestations degrees of COVID-19. For this reason, it represents a good test bed to assess model portability. Exploiting two independent
Figure 9. Results of the proposed system on five examples from the test-set. (top) Three good predictions. (bottom) Two cases chosen between the worst prediction regarding the original clinical annotation $H$. For each block, the most left image is the input CXR that the network analyses, followed by the aligned and masked lungs, to analyse the quality of the segmentation and alignment block. While in the second row we show the Predicted Brixia-score with the clinical ground truth $H$, and the explainability maps. In those maps the relevance goes from white colour (i.e., no contribution to that prediction) to the class colour (i.e., the region had an important weight for the class decision).

Brixia-score annotations of this dataset (one from a senior, and another from a junior radiologist) we perform a portability study to derive some useful guidance for extended use of our model on data generated in other facilities. To this aim we carried out three tests involving only the BS-Net-HA configuration: 1) we directly measured the network performance on the whole set of 194 images considering it a new test set (T-full); and, after random 75/25 splitting of the dataset we performed a performance evaluation on the so reduced test set (T-red), of the BS-Net-HA after both 2) fine-tuning and 3) retraining (starting from both segmentation and alignment block trained on their specific datasets). We considered the senior radiologist as the reference in order to have the possibility to assess the second rater performance the same way we assess the network performance. Table 9 lists all results from the above described tests.

Looking at results on T-full, we can derive that, even by directly applying the model trained on the Brixia-dataset on a completely different, and more variegated dataset (an aggregation of CXR images collected in several centers worldwide, at various spatial resolutions, and other unknown image quality parameters, such as modality and window-level settings), the network holds up under this stress, confirming the meaningfulness of the learning task and already a fair robustness to work in different context even in an uncontrolled way. On the other hand, the skilled human observer confirms higher generalization capability. On the same reduced dataset, when retrained from scratch, the network is not able anymore to produce even the results obtained by the same model trained on the Brixia-dataset. This is a clear evidence of the need
Table 9
Portability tests on the public dataset. MAE and its SD are listed for both reporting radiologist and BS-Net-HA. The network has been used in three training conditions: 1) as is (originally trained on the Brixia-dataset), 2) completely retrained (classification part) on the public dataset, and 3) fine-tuned on the public dataset.

| Test         | A  | B  | C  | D  | E  | F  | Avg. on regions | Global score |
|--------------|----|----|----|----|----|----|----------------|--------------|
| Radiologist  |    |    |    |    |    |    | 0.131          | 0.786        |
| subset       | 0.149 | 0.234 | 0.191 | 0.000 | 0.191 | 0.021 | 0.131          | 0.787        |
| BS-Net-HA    |    |    |    |    |    |    | -0.143         | -0.859       |
| * from scratch | 0.125 | -0.104 | -0.188 | 0.167 | -0.063 | -0.583 | -0.108         | -0.646       |
| * fine-tuning |    |    |    |    |    |    | 0.177          | 1.063        |
| Radiologist  |    |    |    |    |    |    | 0.405          | 1.974        |
| subset       | 0.404 | 0.447 | 0.489 | 0.340 | 0.362 | 0.447 | 0.415          | 1.851        |
| BS-Net-HA    |    |    |    |    |    |    | 0.518          | 2.214        |
| * from scratch | 0.458 | 0.479 | 0.521 | 0.375 | 0.479 | 0.625 | 0.490          | 2.396        |
| * fine-tuning |    |    |    |    |    |    | 0.517          | 2.188        |
| Radiologist  |    |    |    |    |    |    | 0.465          | 2.000        |
| subset       | 0.375 | 0.646 | 0.604 | 0.458 | 0.542 | 0.479 | 0.465          | 2.000        |

7. Final remarks and future directions

We have introduced an end-to-end image analysis platform for the estimation of a semiquantitative rating based on a difficult visual task. Such lung severity score is, by itself, the result of a compromise: on the one hand, the need for a clinically expressive granularity of the different stages of the disease; on the other hand, the built-in subjectivity in the interpretation of CRX images, stemming from the intrinsic limits of such imaging modality, and from the high variability of COVID-19 manifestations. As an additional complication, the available Brixia-score, even if coming from expert personnel, has neither ground truth characteristics (as ratings are affected by inter-observer variability), nor it is highly accurate in terms of spatial indication (since scores are related to generic rectangular regions).

Despite such difficulties, our study is justified and driven by the strong clinical reasons related to the role of CXR images for the management of COVID-19 patient monitoring, especially in conditions of overloaded healthcare facilities. Working with a very large dataset of almost 5000 images allowed to develop a solution which deals with all aspects of the problem, also thanks to other datasets which are exploited in a "from part to whole" training process. The prospected solution is designed to work in a weakly supervised context and to be capable of manifesting self-attentive abilities. Having tested architectural variants, and targeted ablation studies, the network performance ultimately surpasses qualified radiologists in rating accuracy and consistency. We also collected evidences of the robustness and portability of the model (of which we release the code for research use), that favorably highlights its application to other clinical settings.

The automated AI-driven scoring is not meant to eliminate the evaluation of the radiologist, but to aid the reporting workflow, and improving the timeliness of first evaluation, by proposing a preliminary interpretation of findings. Such aid might also prove useful in case of shortage...
of radiologists, condition that might be worsened due to the spread of contagion among physicians.

The main limitations of this work are related to residual errors. Despite highly aligned with radiologists’ performance and not evidencing statistical biases, a case-by-case deeper comparison of possible causes of error could guide further improvement of the image analysis architecture. In particular, since the segmentation task is sometimes ambiguous and hard to accomplish, better solutions should be designed to handle anomalies in the segmentation results.

Regarding future work, an interesting clinical scenario to evaluate is the following: whether, in case the BS-Net follows the same patient, it could exhibits greater self-coherent behavior if compared to the case where serial CXR acquisitions are reported by different radiologists (according to availability and working shifts). Another direction to consider is the possibility to exploit CT images from COVID-19 patients to derive semiquantitative ground truth information directly from the quantitative volumetric assessments (as in (Gozes et al., 2020)). Last, it could be relevant to check the portability of this architecture in contexts where the scoring system is different, such as in cases where the score is defined with lower granularity (Toussie et al., 2020).

Declaration of Competing Interest
None.

Acknowledgements
The authors would like to thank colleagues of the Information and IT System Operational Units of ASST Spedali Civili of Brescia, Francesco Scazzoli, Silvio Finardi, Roberto Marini, and Sabrina Vicari for providing the infrastructure. A very special thanks to Marco Renato Roberto Merli, and Andrea Carola of Philips S.p.A., and to Gian Stefano Bosio, and Simone Gaggero of EL.CO. S.r.l (Caïro Montenotte, Italy) for their outstanding technical support.

References
Bontempi, D., Benini, S., Signoroni, A., Svanera, M., Muckli, L., 2020. CEREBRUM: a fast and fully-volumetric Convolutional Encoder-decoder for weakly-supervised Segmentation of Brain structures from out-of-the-scanner MRI. Medical Image Analysis 62, 101688.

Borghesi, A., Maroldi, R., 2020. COVID-19 outbreak in Italy: experimental chest X-ray scoring system for quantifying and monitoring disease progression. La radiologia medica.

Borghesi, A., Zigliani, A., Golemi, S., Carapella, N., Maculotti, P., Farina, D., Maroldi, R., 2020a. Chest X-ray severity index as a predictor of in-hospital mortality in coronavirus disease 2019: A study of 302 patients from Italy. International Journal of Infectious Diseases 96, 291 – 293.

Borghesi, A., Zigliani, A., Masciiullo, R., Golemi, S., Maculotti, P., Farina, D., Maroldi, R., 2020b. Radiographic severity index in COVID-19 pneumonia: relationship to age and sex in 783 Italian patients. La radiologia medica.

Burlacu, A., Crisan-Dabija, R., Popa, I.V., Artene, B., Birzu, V., Pricop, M., Plesoiianu, C., Generali, D., 2020. Curbing the AI-induced enthusiasm in diagnosing COVID-19 on chest X-Rays: the present and the near-future. medRxiv.

Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A., 2020. Albumentations: Fast and flexible image augmentations. Information 11.

Candemir, S., Antani, S., 2019. A review on lung boundary detection in chest x-rays. International Journal of Computer Assisted Radiology and Surgery 14, 563–576.

Candemir, S., Jaeger, S., Palaniappan, K., Musco, J.P., Singh, R.K., Xue, Z., Karargyris, A., Antani, S., Thoma, G., McDonald, C.J., 2014. Lung Segmentation in Chest Radiographs Using Anatomical Atlases With Nonrigid Registration. IEEE Transactions on Medical Imaging 33, 577–590.
D., Messa, C., Sardanelli, F., 2020. Artificial intelligence applied on chest X-ray can aid in the diagnosis of COVID-19 infection: a first experience from Lombardy, Italy. medRxiv.

Cohen, J.P., Morrison, P., Dao, L., 2020. COVID-19 image data collection. arXiv 2003.11597 URL: https://github.com/ieee8023/covid-chestxray-dataset.

Filice, R.W., Ratwani, R.M., 2020. The Case for User-Centered Artificial Intelligence in Radiology. Radiology: Artificial Intelligence 2, e190095.

Frid-Adar, M., Ben-Cohen, A., Amer, R., Greenspan, H., 2018. Improving the segmentation of anatomical structures in chest radiographs using u-net with an imagenet pre-trained encoder, in: Image Analysis for Moving Organ, Breast, and Thoracic Images. Springer, pp. 159–168.

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 and patient monitoring using deep learning CT image analysis. arXiv:2003.05037.

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.

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.

Huang, L., Han, R., Ai, T., Yu, P., Kang, H., Tao, Q., Xia, L., 2020. Serial quantitative chest ct assessment of COVID-19: Deep-learning approach. Radiology: Cardiothoracic Imaging 2, e200075.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Hagghoo, B., Ball, R., Shpanskaya, K., et al., 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 590–597.

Jaderberg, M., Simonyan, K., Zisserman, A., kavukcuoglu, k., 2015. Spatial Transformer Networks, in: Advances in Neural Information Processing Systems 28. Curran Associates, Inc., pp. 2017–2025.

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. Quantitative imaging in medicine and surgery 4, 475.

Kalkreuth, R., Kaufmann, P., 2020. COVID-19: A survey on public medical imaging data resources. arXiv:2004.04569.

Karim, M.R., Dhmen, T., Rebholz-Schuhmann, D., Decker, S., Cochez, M., Bevan, O., 2020. DeepCOVIDExplainer: Explainable COVID-19 predictions based on chest x-ray images. arXiv:2004.04582.

Karimi, D., Dou, H., Warfield, S.K., Gholipour, A., 2019. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. arXiv:1912.02911.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kundu, S., Elhalawani, H., Gichoya, J.W., Kahn, C.E., 2020. How might ai and chest imaging help unravel COVID-19’s mysteries? Radiology: Artificial Intelligence 2, e200053.

Laghi, A., 2020. Cautions about radiologic diagnosis of COVID-19 infection driven by artificial intelligence. The Lancet Digital Health 2, e225.

Latif, S., Usman, M., Manzoor, S., Iqbal, W., Qadir, J., Tyson, G., Castro, I., Razi, A., Boulos, M.N.K., Weller, A., Crowcroft, J., 2020. Leveraging Data Science To Combat COVID-19: A Comprehensive Review.

Li, X., Li, C., Zhu, D., 2020. COVID-MobileXpert: On-device COVID-19 screening using snapshots of chest x-ray. arXiv:2004.03042.

Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature Pyramid Networks for Object Detection, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
Linda Wang, Z.Q.L., Wong, A., 2020. COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images. arXiv:2003.09871.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440.

Maguolo, G., Nanni, L., 2020. A Critic Evaluation of Methods for COVID-19 Automatic Detection from X-Ray Images. arXiv:2004.12823.

Manna, S., Wruble, J., Maron, S., Toussie, D., Voutsinas, N., Finkelstein, M., Cedillo, M.A., Diamond, J., Eber, C., Jacobi, A., Chung, M., Bernheim, A., 2020. COVID-19: A Multimodality Review of Radiologic Techniques, Clinical Utility, and Imaging Features. Radiology: Cardiothoracic Imaging 2, e200210.

Murphy, K., Smits, H., Knoops, A.J., Korst, M.B., Samson, T., Scholten, E.T., Schalekamp, S., Schaefer-Prokop, C.M., Philipson, R.H.H., Meijers, A., Melendez, J., van Ginneken, B., Rutten, M., 2020. COVID-19 on the chest radiograph: A multi-reader evaluation of an ai system. Radiology 0, 201874.

Oh, Y., Park, S., Ye, J.C., 2020. Deep learning COVID-19 features on cxr using limited training data sets. IEEE Transactions on Medical Imaging , 1–1.

Pereira, R.M., Bertolini, D., Teixeira, L.O., Silla, C.N., Costa, Y.M., 2020. COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios. Computer Methods and Programs in Biomedicine , 105532.

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. arXiv:2004.08379.

Ramachandran, P., Zoph, B., Le, Q.V., 2017. Searching for activation functions. arXiv preprint arXiv:1710.05941.

Reyes, M., Meier, R., Pereira, S., Silva, C.A., Dahlweid, F.M., Tengg-Kobligk, H.v., Summers, R.M., Wiest, R., 2020. On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. Radiology: Artificial Intelligence 2, e190043.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. Why Should I Trust You?: Explaining the Predictions of Any Classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA. p. 11351144.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Rubin, G.D., Ryerson, C.J., Haramati, L.B., Sverzellati, N., Kanne, J.P., Raoof, S., Schwager, N.W., Volpi, A., Yim, J.J., Martin, I.B.K., Anderson, D.J., Kong, C., Altes, T., Bush, A., Desai, S.R., Goldin, J., Goo, J.M., Humbert, M., Inoue, Y., Kauczor, H.U., Luo, F., Mazzone, P.J., Prokop, M., Remy-Jardin, M., Richeldi, L., Schaefer-Prokop, C.M., Tomiyama, N., Wells, A.U., Leung, A.N., 2020. The Role of Chest Imaging in Patient Management during the COVID-19 Pandemic: A Multinational Consensus Statement from the Fleischner Society. Radiology (simultaneously published in Chest) 0, 201365.

Sardanelli, F., Di Leo, G., 2020. Assessing the Value of Diagnostic Tests in the New World of COVID-19 Pandemic. Radiology X, 201845.

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: The IEEE International Conference on Computer Vision (ICCV).

Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., He, K., Shi, Y., Shen, D., 2020. Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation and Diagnosis for COVID-19. IEEE Reviews in Biomedical Engineering , 1–1.

Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K.i., Matsui, M., Fujita, H., 21
Kodera, Y., Doi, K., 2000. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists’ detection of pulmonary nodules. American Journal of Roentgenology 174, 71–74.

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

SIRM, 2020. Utilizzo della diagnostica per immagini nei pazienti COVID-19. Italian Society of Medical Radiology. URL: https://www.sirm.org.

Stirenko, S., Kochura, Y., Arienin, O., Rokovyi, O., Gordienko, Y., Gan, P., Zeng, W., 2018. Chest x-ray analysis of tuberculosis by deep learning with segmentation and augmentation, in: 2018 IEEE 38th International Conference on Electronics and Nanotechnology (ELNANO), IEEE. pp. 422–428.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-first AAAI conference on artificial intelligence.

Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X., 2020. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. Medical Image Analysis 63, 101693.

Tan, M., Pang, R., Le, Q.V., 2019. EfficientDet: Scalable and Efficient Object Detection. arXiv:1911.09070.

Tartaglione, E., Barbano, C.A., Berzovini, C., Calandri, M., Grangetto, M., 2020. Unveiling COVID-19 from chest x-ray with deep learning: a hurdles race with small data. arXiv:2004.05405.

Ting, D.S.W., Carin, L., Dzau, V., Wong, T.Y., 2020. Digital technology and COVID-19. Nature Medicine 26, 459–461.

Toussie, D., Voutsinas, N., Finkelstein, M., Cedillo, M.A., Manna, S., Maron, S.Z., Jacobi, A., Chung, M., Berheim, A., Eber, C., Concepcion, J., Fayad, Z., Gupta, Y.S., 2020. Clinical and chest radiography features determine patient outcomes in young and middle age adults with COVID-19. Radiology 0, 201754.

Vedaldi, A., Soatto, S., 2008. Quick shift and kernel methods for mode seeking, in: European conference on computer vision, Springer. pp. 705–718.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3462–3471.

WHO, 2020. Coronavirus disease (COVID-19) outbreak. World Health Organization. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

Willemink, M.J., Koszek, W.A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L.R., Summers, R.M., Rubin, D.L., Lungren, M.P., 2020. Preparing Medical Imaging Data for Machine Learning. Radiology 295, 4–15.

Xu, Y., Zhu, J.Y., Chang, E.I.C., Lai, M., Tu, Z., 2014. Weakly supervised histopathology cancer image segmentation and classification. Medical Image Analysis 18, 591 – 604.

Zhou, Z., Rahman Siddiquee, M.M., 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 International Publishing, Cham. pp. 3–11.

Zhou, Z.H., 2017. A brief introduction to weakly supervised learning. National Science Review 5, 44–53.