Chapter 1
AI and Point Of Care Image Analysis For COVID-19

Michael Roberts, Oz Frank, Shai Bagon, Yonina C. Eldar, and Carola-Bibiane Schönlieb

Abstract Point-of-care imaging, including chest x-ray, computed tomography and ultrasound have played a critical role in the response to COVID-19. Clinicians study the patterns and changes in the images to make a likely diagnosis, assess a patients clinical response and to predict likely outcomes. In this chapter we focus on the application of artificial intelligence (AI) techniques to these imaging modalities and discuss the contributions in the literature for diagnosis and prognostic models.

1.1 Introduction

Point of care image analysis for hospitalised patients remains a largely manual process. Typically, in Europe and the US, when a patient suspected to have COVID-19 is admitted to the hospital, a Chest X-Ray (CXR) is acquired and a reverse

Michael Roberts
Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK. e-mail: mr808@cam.ac.uk
Oncology R&D, AstraZeneca, Cambridge, UK e-mail: michael.roberts2@astrazeneca.com

Oz Frank
Weizmann AI Center (WAIC), Faculty of Mathematics and Computer Science, Weizmann Institute of Science, Israel. e-mail: oz.frank@weizmann.ac.il

Shai Bagon
Weizmann AI Center (WAIC), Faculty of Mathematics and Computer Science, Weizmann Institute of Science, Israel. e-mail: shai.bagon@weizmann.ac.il

Yonina C. Eldar
Weizmann AI Center (WAIC), Faculty of Mathematics and Computer Science, Weizmann Institute of Science, Israel. e-mail: yonina.eldar@weizmann.ac.il

Carola-Bibiane Schönlieb
Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK. e-mail: cbs31@cam.ac.uk

transcription polymerase chain reaction (RT-PCR) test is performed. This CXR is then interpreted by radiologists and other clinicians who will identify particular patterns which may indicate a likely initial diagnosis and prognosis. When the results of the RT-PCR test are returned, it will replace, or confirm, this initial image-based diagnosis. If the CXR or the patient display any particular complications then further imaging, such as Computed Tomography (CT) or Ultrasound (US) can be requested, which are then interpreted by clinical staff once more.

In a pandemic, when every resource in the hospital system is under strain, it is imperative to make manual processes as efficient as possible to reduce the time for decisions to be taken and actions made. Artificial Intelligence (AI) algorithms and methods promise great potential for automating many routine tasks for clinicians and thereby the promise to improve clinical care. This includes, but is not limited to (a) identifying regions of pathology on images, (b) tracking disease burden in longitudinal imaging and (c) measuring regions and volumes of interest.

As of 2021, we remain a long way from point of care imaging that is informed by AI techniques being available routinely – although it feels tantalisingly close to being a reality. The COVID-19 pandemic has highlighted how AI-based algorithms could have a significant impact – if only they were available to call upon during the height of the pandemic. In particular, if COVID-19 diagnosis were possible based on e.g. an admission CXR, it would have been possible to triage patients immediately to the “green" wards" of non-COVID-19 patients and the “red" wards of COVID-19 cases. Unfortunately, when RT-PCR testing capacity was restricted, clinicians would wait anything from 24-48 hours for a positive or negative result to be returned. This led to inevitable cross-infection and even more extreme stress on the health system.

Similarly, prognostication for COVID-19 patients using AI-based methods has the promise to provide huge improvement in clinical care and allow for better resource management. Examples of prognostication tasks are: (a) prediction of ventilation requirement and the level of ventilation required, (b) prediction of response to treatments (such as Dexamethasone) and the ideal time to administer for optimal response, (c) prediction of the patients that will experience acute respiratory distress syndrome (ARDS).

In this Chapter we focus on the reality, rather than the promise, of how AI was applied to point of care imaging (CXR, CT and US) in the COVID-19 pandemic. We will review each imaging modality separately and highlight models described in the literature along with common themes, pitfalls and recommendations for how future models can be developed following best practice. We conclude this chapter by providing some success stories and reasons for optimism, highlight some of the dangers of developing models which do not perform as expected along with lessons learned from this pandemic that we can take forward to be better prepared for the next one.

1.1.1 Motivation for using imaging

As the COVID-19 pandemic swept through China in early 2020, chest imaging was used locally as the primary initial diagnostic tool. Meanwhile, European and American radiological societies did not initially support the use of CT and CXR imaging for diagnosis in early March 2020 [1, 2] with the ACR stating

> CT should not be used to screen for or as a first-line test to diagnose COVID-19

but this position softened towards the end of March as the pandemic took hold and testing capacity was limited with an update in late March 2020 stating

> The ACR strongly urges caution in taking this approach [...] Clearly, locally constrained resources may be a factor in such decision making.

Several studies also indicate that, in addition to imaging being a potential diagnostic tool, it also encodes prognostic information about the disease. For example, the extent of opacification in the lungs of COVID-19 patients is a significant prognostic marker of mortality [3].

Fig. 1.1, courtesy of [6], displays common presentations of COVID-19 in CT scans and CXRs. In both the CXR and CT imaging we see ground-glass opacities in the regions affected by COVID-19, with the CT scans showing a crazy-paving style pattern inside those ground-glass opacities.

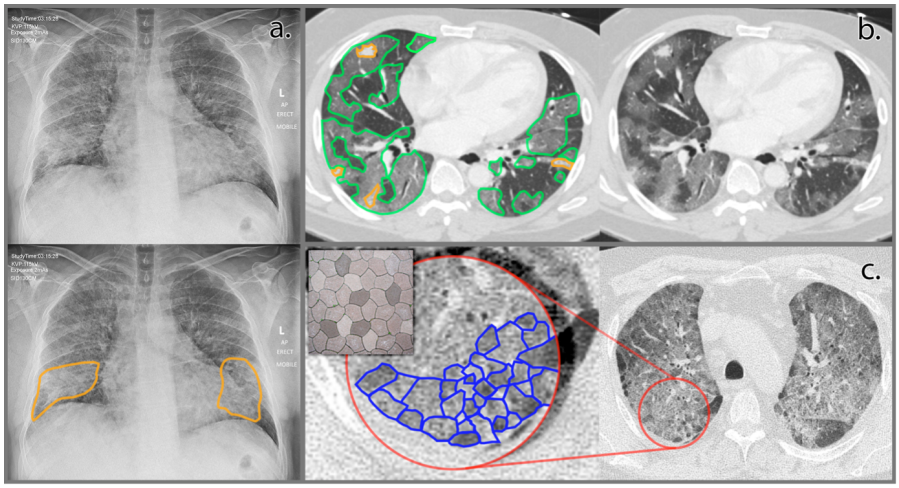


Fig. 1.1 Annotated examples of COVID-19 scans. (a) Chest X-ray (CXR) with ground-glass opacification in both lungs and consolidation (outlined in orange). (b) A CT scan that shows ground-glass opacification (green) and consolidation (orange). (c) A CT scan that indicates severe COVID-19 with a crazy-paving pattern. Images from the NCCID [4], and inset from [5].

1.1.2 Motivation for using AI with imaging

The ground-glass opacification of the lungs with a 'crazy-paving' pattern seen in CT scans for COVID-19 patients motivates the idea that pattern recognition algorithms hold the potential to aid clinicians in the diagnosis and prognostication of COVID-19 via chest imaging [7].

The COVID-19 pandemic is the first of the machine learning era and, given recent developments in the application of machine learning models to medical imaging problems [7, 9, 25, 26], there is fantastic promise for applying machine learning methods to COVID-19 radiological imaging for improving the accuracy of diagnosis, compared to the gold-standard RT-PCR, whilst also providing valuable insight for prognostication of patient outcomes. These models have the potential to exploit the large amount of multi-modal data collected from patients and could, if successful, transform detection, diagnosis, and triage of patients with suspected COVID-19. One model of huge potential utility is a model which can not only distinguish COVID-19 from non-COVID-19 patients but also discern alternative types of pneumonia such as those of bacterial or other viral aetiologies. For prognostication, it would is desirable to develop models that predict responses to therapies and clinical pathways for patients. This would allow for resource forecasting, patient triage and ultimately improved care.

1.1.3 Integration of imaging with other modalities

In developing machine learning models for COVID-19 the model input can be from a variety of sources, including but not limited to electronic health record records, full blood count data, chest imaging, audio recording of coughs, symptom diaries, genetics, and more. A single patient will have data recorded about them in several modalities and at different levels of granularity. We aim to demonstrate why it is important to combine/fuse data from multiple sources to get models that make predictions using holistic understanding of the patient's condition.

In late February 2020, the Diamond Princess cruise ship had the largest cluster of positive COVID-19 cases outside of China. A study of 104 of these COVID-19-positive patients found that 73% (76 out of 104) were asymptomatic. However, 54% (41 out of 76) of these asymptomatic individuals displayed lung opacities on their CT scans. The converse was also true, as roughly 21.5% (6 out of 28) of symptomatic patients had normal CT findings [10]. Imaging features alone are clearly not sufficient for accurate diagnosis and neither are the clinical features alone enough to understand the degree of disease in a patient.

It is also important to recognise that clinicians routinely use multiple data sources to develop their judgment for a patient's likely diagnosis and their likely clinical outcome. However, although it is simple for the clinician to do this, fusing multi-modal data is not trivial in a machine learning framework.

1.1.4 Literature Overview

There are a huge number of papers which discuss machine learning models for COVID-19 diagnosis or prognosis using point of care imaging. Indeed, for 2020 and 2021, a basic search of Arxiv, BioRxiv, MedRxiv and Pubmed for papers which mention machine/deep learning and COVID-19 and CT/CXR/US imaging returns 848 results with 294 of these being preprints. The three journals publishing the most papers on the subject were Scientific Reports (17), IEEE Journal of Biomedical and Health Informatics (16) and PLoS ONE (14).

In keeping with this large corpus of literature and the fast moving nature of the pandemic, there have also been many systematic reviews, 27 in total with 16 of them published. The systematic reviews most relevant to this Chapter are [13, 17, 14, 15, 12, 16, 8, 11, 18, 19, 20, 21, 22, 23, 24, 28, 29, 30, 31, 32, 33, 34, 35].

Several of the authors of this chapter performed the systematic review [8] which examines the entire literature from January 1, 2020 to October 3, 2020 and identifies 320 published and preprint manuscripts that develop machine learning models using chest CT or radiographs for COVID-19 diagnosis or prognostication. Unfortunately, as we will discuss throughout this Chapter, many of the papers contained systematic issues pertaining to image sourcing, quality, and documentation that introduce bias in developed models, ultimately making them unlikely to perform well in practice [6].

In our search of the literature, we found that 637/848 (75.1%) of papers mention diagnosis, detection, diagnostics, screening, recognition, discrimination, identification or classification of COVID-19 in the title of the manuscript whilst only 94/848 (11.1%) consider prognostication (17 papers consider both). For completion, the remaining papers are 59/848 (7.0%) which design segmentation models for COVID-19 patterns in imaging (15 of which also perform diagnosis) and the remaining 82 papers are literature reviews, introduce new COVID-19 datasets or discuss methodologies such as image reconstruction or denoising.

This significant bias towards the development of diagnosis models is understandable as at the start of the pandemic there was a hunger for a solution to the slow processing of RT-PCR tests globally (typically 24-48 hours) which would allow for rapid triage of patients into isolation if needed. COVID-19 imaging and clinical data was also scarce and precious at the start of the pandemic, where non-COVID-19 data was relatively plentiful, and it was therefore easier to develop methods which aim to detect the COVID-19 data among a sea of other data. Prognostic models require well curated data, with imaging linked to standardised outcomes which is harder to collect. This is a challenge with AI methods, which rely on large, diverse data sets, in order to succeed.

While Xray and CT are commonly used during the treatment of respiratory conditions, lung ultrasound (LUS) is not an obvious choice due to the unique acoustical properties of the lungs. We argue that despite its challenging properties LUS can and should be considered as a point-of-care modality, and further show how AI can assist in achieving this goal. In the rest of the chapter, we discuss AI methods for Xray, CT and ultrasound imaging, respectively.

1.2 Chest X-Ray Imaging

In Europe, the US and most countries of the world chest X-ray imaging is used as the first-line imaging given to any COVID-19 patient once they enter the hospital. Chest X-rays are fast to acquire and relatively low cost, so a patient is highly likely to have several in a normal admission. Early into the pandemic, when RT-PCR results were commonly only returned 24 hours or more after the sample was taken, the CXR images were used in combination with clincal symptoms to make a likely diagnosis of the patients. In addition to the initial diagnosis of patients, CXRs can be used to determine a prognosis for the patient based on location and extent of the ground-glass opacities and consolidation in the lungs. In this section, we discuss many of the approaches taken to diagnosis and prognosis of COVID-19 using CXR images along with specific focus on those which use longitudinal imaging in their models and fuse data from non-imaging modalities into their input features. We conclude by highlighting many of the issues that are common across the literature.

1.2.1 Diagnosis Models

Machine learning models for the diagnosis of COVID-19 using CXR imaging and tend to pose the problem as either a two class problem of COVID-19 vs. Non-COVID-19 or a more complex multi-class problem, such as COVID-19 vs. bacterial pneumonia vs. other viral pneumonia vs. healthy. The latter is more nuanced as it requires careful definition and identification of the non-COVID-19 classes to which the model is likely to be applied in practice whereas a two-class model collates all the non-COVID-19 data into one class. Most papers classify images into the three classes COVID-19, non-COVID-19 pneumonia and normal while several consider an extra class by dividing non-COVID-19 pneumonia into viral and bacterial pneumonia.

Commonly, the CXR images are pre-processed by segmenting the lungs to remove biases in the images, such as labels imprinted on the image, and artefacts around the border of the CXR. In [27], the authors consider a segmentation path input together with the image itself, and show that this enhances performance. They also introduce further pre-processing techniques based on augmentations inspired by clinical input which further improves detection performance.

Most papers directly apply deep learning methodologies to the question of diagnosis. Simply using the images and associated labels to train networks which learn their own features. For this, most authors chose to use transfer learning, taking models trained on existing CXR datasets such as CheXpert and fine-tuning them to the COVID-19 data. Most papers use off-the-shelf network architectures based on the root model being used for transfer learning, including ResNet-18 or ResNet-50, DenseNet-121, VGG-16 or VGG-19, Inception and EfficientNet. Few papers develop their own custom architectures. ResNet and DenseNet architectures reported better performance than the others, with accuracies ranging from 0·88 to 0·99. However, we caution against direct comparison since the papers use different training and test-

ing settings (e.g. different datasets and data partition sizes) and consider a different number of classes.

Some papers use a more traditional approach, extracting hand engineered features from the images and linking these to the outcomes by fitting a machine learning model, such as a random forest or neural network.

1.2.2 Prognosis Models

As discussed, the literature for machine learning algorithms that prognosticate for COVID-19 patients is small compared to diagnostic algorithms. This is due to difficulty in identifying, anonymising and extracting datasets of images and clinical outcomes which are linked to one another and of high quality.

Outcomes which are typically predicted by these models are those such as: death or need for ventilation, a need for ICU admission, progression to acute respiratory distress syndrome, the length of hospital stay, likelihood of conversion to severe disease and the extent of lung infection.

Predictors from radiological data were extracted using either handcrafted radiomic features or deep learning. Clinical data included basic observations, serology and comorbidities. Most papers used models based on a multivariate Cox proportional hazards model, logistic regression, linear regression, random forest or compare a huge variety of machine learning models such as tree-based methods, support vector machines, neural networks and nearest neighbour clustering.

1.2.3 Use of longitudinal imaging

As CXR imaging is the most common imaging modality for POC in Europe, the USA and most countries of the world, it is common for patients to have multiple images captured in a single hospital stay. If a model is able to read in several images in sequence for a patient then it is possible to capture the changes in features of the disease as it progresses and this should allow for more accurate prediction of their outcomes. This approach introduces some methodological challenges, such as: images acquired on different machines and of different quality, images acquired at non-equidistant intervals and that the baseline images for all patients will be acquired at different stages of their disease depending on when they presented to hospital.

In the literature, there are several examples of papers which consider longitudinal CXR imaging for COVID-19 prognostication, in particular [36, 37]. These papers employ different techniques to fuse the features extracted from the longitudinal images. In [36], the authors use a self-supervised approach using momentum and contrastive learning (MoCo) [XXX] to train a feature extractor for the CXR images. This feature extractor is applied to each image in the sequence and the relative time of the image acquistion is concatenated to the resulting feature vector. Therefore, for

each image we have one feature vector which encodes the image features and the relative timing. After this, a Transformer [XXX] is applied to the sequence of feature vectors from all of the images and classification is made for whether an adverse event (death, ICU admission or intubation) takes place within 24, 48, 72 or 96 hours of the CXR.

In [37], the authors use a Cox proportional hazards model to predict the probability of experiencing death, ICU admission, ICU discharge, hospital admission and hospital discharge before the observing time t. They extract features from the images using a convolutional LSTM, concatenate them and use fully connected layers to obtain a risk score.

The literature consistently suggests that using multiple CXR images gives better model performance than using a single time point.

1.2.4 Fusion with other data modalities

Very few papers discuss models which integrate both radiological and clinical data for COVID-19 diagnosis and prognostication. Fusing data of these different modalities is a particular challenge, as these data are of different types and dimensionalities. In particular, clinical data tend to be vectors or tables of features for each patient, potentially with multiple time points per patient. Imaging is not only extremely high-dimensional, with millions of pixels in a normal CXR, but there are also spatial relationships between pixels and structures in the image to consider. Methodologies for fusing this data of different modalities (which is also potentially longitudinal) are being actively studied in the literature [39].

There are some examples in the literature, such as [41, 38], which combine CXR imaging features with clinical data in a machine learning model to predict outcomes for COVID-19 patients. In both cases, these papers state that incorporation of clinical data, in addition to imaging features, improved the performance of the model. The papers take very different approaches to fusing the different modalities.

In [41], the authors manually graded severity of the disease in different zones of the CXR and use these hand-engineered features in combination with the clinical data in a variety of machine learning models (support vector machines, random forest, linear regression and XGBoost). They find that the model developed using a combination of the CXR and clinical features performs better than those developed using only the CXR and clinical features. In [38], the authors use a convolutional neural network to extract the features from the CXR imaging and fuse the clinical data into the final fully connected layer by concatenation. They similarly find that a model developed using the fusion of imaging and clinical data is better than those developed on the modalities separately.

1.2.5 Common issues with AI and Chest X-Ray imaging

With no standardisation, AI algorithms for COVID-19 have been developed with a very broad range of applications, data collection procedures and performance assessment metrics. This creates various challenges such as (i) bias in small data sets; (ii) variability of large internationally-sourced data sets; (iii) poor integration of multi-stream data, particularly imaging data; (iv) difficulty of the task of prognostication, and (v) necessity for clinicians and data analysts to work side-by-side to ensure the developed AI algorithms are clinically relevant and implementable into routine clinical care. Since the pandemic began in early 2020, researchers have answered the 'call to arms' and numerous machine learning models for diagnosis and prognosis of COVID-19 using radiological imaging have been developed and hundreds of manuscripts have been written.

In this section, we highlight some of the key systemic issues found in [8] and other systematic reviews.

Duplication and quality issues.

Many papers rely on public COVID-19 datasets, such as [42, 43, 44, 45, 46]. However, there is no restriction for a contributor to upload COVID-19 images to many of these public repositories. There is high likelihood of duplication of images across these sources and no assurance that the cases included in these datasets are confirmed COVID-19 cases (authors take a great leap to assume this is true) so great care must be taken when combining datasets from different public repositories. Also, most of the images have been pre-processed and compressed into non-DICOM formats leading to a loss in quality and a lack of consistency/comparability.

Source issues.

In the literature, many papers use the pneumonia dataset of Kermany et al. [9] as a control (i.e. non-COVID-19) group. However, they commonly fail to mention that this consists of paediatric patients aged between one and five. Developing a model using adult COVID-19 patients and very young pneumonia patients is likely to overperform as it is merely detecting children vs. adults. This dataset is also erroneously referred to as the Mooney dataset in many papers (being the Kermany dataset deployed on Kaggle [47]).

Another significant source issue identified in the systematic reviews of the literature is that data of different classes tend to come from different sources, e.g. RSNA [48] contains only non-COVID-19 pneumonia CXRs, Kermany [9] contains paediatric non-COVID-19 pneumonia CXRs and CheXpert [49] contains CXRs for a range of non-COVID-19 lung diseases. The issue here arises when a machine learning model learns the source of the data rather than imaging features unique to the diseases of interest.

It is demonstrated by Maguolo et al. [50] that by excluding the lung region entirely, the authors could identify the source of the images in the Cohen et al. [42] and Kermany et al. [9] datasets with an AUC between 0.9210 to 0.9997 and 'diagnose' COVID-19 with an AUC=0.68.

Frankenstein datasets.

The previously discussed issues of duplication and sourcing of data become compounded when public 'Frankenstein' datasets are used. These are datasets assembled from other datasets and redistributed under a new name. For instance, one dataset [51] combines several other datasets [44, 42, 9] without realising that one of the component datasets [42] already contains another component [44]. This repackaging of datasets, although pragmatic, inevitably leads to problems with authors developing algorithms (in good faith) which are being trained and tested on identical or overlapping datasets which they believed to be from distinct sources.

Implicit biases in the source data.

Images uploaded to a public repository and those extracted from publications [42] are likely to have implicit biases due to the contribution source. For example, it is likely that more interesting, unusual or severe cases of COVID-19 appear in publications.

The urgency of the pandemic led to many studies using datasets that contain obvious biases or are not representative of the target population, e.g. paediatric patients. Before evaluating a model, it is crucial that authors report the demographic statistics for their datasets, including age and sex distributions. Diagnostic studies commonly compare their models' performance to that of RT-PCR. However, as the ground-truth labels are often determined by RT-PCR, there is no way to measure whether a model outperforms RT-PCR from accuracy, sensitivity, or specificity metrics alone. Ideally, models should aim to match clinicians using all available clinical and radiomic data, or to aid them in decision making.

Artificial limitations due to transfer learning.

Many papers utilise transfer learning in developing their model, which assumes an inherent benefit to performance. However, it is unclear whether transfer learning offers significant performance benefit due to the over-parametrisation of the models [52]. Many publications used the same resolutions such as 224-by-224 or 256-by-256 for training, which are often used for ImageNet classification, indicating that the pre-trained model dictated the image rescaling used rather than clinical judgement. This is particularly hard to justify, given that the features which differentiate COVID-19 pnuemonia from other diseases are likely subtle and not appreciable at such coarse resolutions.

1.3 Computed Tomography Imaging

Computed Tomography (CT) images are far more resource intensive to acquire and analyse than CXR imaging. This is due to the high relative cost of machines, requirement for cleaning between patients, technicians to maintain the equipment and the high level of skill required for their analysis. The acquisition of a scan also exposes a patient to a high dose of radiation. Therefore, in many parts of the world, CT scans were reserved for the most complex clinical COVID-19 cases after an initial CXR. However, in China and Russia, CT was used as a first line imaging modality for COVID-19 patients.

There are fewer papers in the literature which focus on CT imaging, however there are a sizeable number. Just as the pandemic was starting to take hold in Europe in March 2020, its effects were starting to ease in China. As a consequence, many of the earliest papers were describing models developed using Chinese data and focused specifically on CT imaging.

1.3.1 Diagnosis Models

Diagnosis of COVID-19 from a CT scan is most relevant and useful for those countries in which CT scans are used as a first-line imaging modality. In countries which use CXR as first-line, a CT scan is used to assess more complex clinical disease but less for diagnosis of disease. Countries which use CT as a first-line modality will also have more CT scans for COVID-19 cases of mild disease, when it is most useful to accurately diagnose COVID-19 to allow for triage and isolation.

The majority of the papers in the literature apply deep learning directly to the CT diagnosis problem and frame it as a classification task, distinguishing COVID-19 from other lung pathologies such as (viral or bacterial) pneumonia, interstitial lung disease and/or a non-COVID-19 class.

In most of these papers, authors consider isolated 2D slices or even 2D patches taken from the 3D volume. This is usually due to computational and storage constraints, as each CT image is typically between 200MB and 500MB and tens or hundreds of millions of voxels. In most 2D models, authors employed transfer learning, with networks pre-trained on ImageNet [53]. Almost all models used lung segmentation as a pre-processing step.

Recognising that there was a relatively small cohort of COVID-19 CT scan data available early in the pandemic, some authors e.g. [55] use Generative Adversarial Network (GAN) [54] approach to create synthetic COVID-19 imaging.

Outside of deep learning based models, other papers in the literature considered more traditional machine learning methods for COVID-19 diagnosis relying on hand-engineered features or CNN-extracted features. Software such as PyRadiomics [XXX] was commonly used to extract a large number of radiomic features from delineated regions in the CT scans and, after feature reduction, a classifier was fit to the remaining features, with most authors using logistic regression.

The performance of models in the literature is highly variable and optimistic with many reporting AUC, sensitivity and specificity values over 0.95. For reference, the gold standard RT-PCR test has a sensitivity of around 80% [XXX].

1.3.2 Prognosis Models

The literature for prognostic models using CT imaging, as with CXR imaging, is relatively small compared to the diagnostic models but approaches have been developed using similar models and features. There is a bias for prognostic models to focus on CT imaging rather than CXR.

As for CXR, models were developed for predicting severity of outcomes including: death or need for ventilation, a need for ICU admission, progression to acute respiratory distress syndrome, the length of hospital stay, likelihood of conversion to severe disease and the extent of lung infection. The features are either handcrafted radiomic features or learned features extracted from a CNN. Most papers fit models based on a multivariate Cox proportional hazards model, logistic regression, linear regression, random forest or compare a huge variety of machine learning models such as tree-based methods, support vector machines, neural networks and nearest neighbour clustering.

1.3.3 Applications to regions away from the lungs

In the literature, almost all papers for CT prognosis and diagnosis of COVID-19 focus on either the full CT scan or the segmented lungs and use these as inputs to the models. However, this ignores many of the other important structures of the body which can be captured in a CT scan.

Most importantly, as poor cardiovascular health one of the largest risk factors for a poor outcome for COVID-19 patients [56], it is highly relevant to consider the heart features in CT images for prognostic models. For example, identifying and quantifying atherosclerosis, tracking changes in the features of the heart through the course of disease and quantifying the extent and volume of epicardial adipose tissue (EAT), i.e. fat around the heart.

In [57], for example, the authors use a semi-automated tool to quantify the EAT around the heart tissue in CT scans for COVID-19 patients and find that there is a link between EAT and the burden of the COVID-19 pneumonia in the lung (ground-glass opacities and consolidation).

In [62], the authors find a higher prevalence of pulmonary embolisms (PEs) in COVID-19 patients and in [63] the authors discuss a deep learning algorithm for identifying PEs. Further study of the vascular structure could also be performed along with a study of changes in morphology and how these link to outcomes.

In the literature considering lung diseases and imaging, airway features have been postulated as markers of disease progression [58, 59]. Not only can one consider the diameter changes and total volume features, but the overall morphology of this complex structure can also be considered and changes monitored. There is not currently existing literature exploring this.

Finally, we also mention that as obesity is known to be a risk factor for poor COVID-19 outcomes [56], it is not unreasonable to consider the features of the liver if an abdominal CT is acquired alongside the thoracic CT. Hepatic steatosis is an accumulation of fat inside the liver, which is closely linked to obesity, that is found to be of higher prevalence in COVID-19 patients [60]. It is unexplored whether features of this fatty tissue, and the liver as a whole could harbour prognostic imaging features.

1.3.4 Use of longitudinal imaging

Acquisition of CT imaging requires exposing the patient to a non-trivial amount of radiation and therefore CT imaging for COVID-19 is only recommended when clinically necessary in the case of seriously ill patients [64] or where the patient condition is worse than would be expected based on the CXR. In our search of the literature, no papers were found which discussed using longitudinal CT imaging and machine learning models. This is potentially due to the fact that the added value of using multiple CT scans to monitor COVID-19 pneumonia has been questioned [61, 65] and found to be of little contribution. Therefore, CTs are not used for routine disease monitoring, and the population will be highly biased. This contrasts with the use of CXR for routine disease monitoring.

Therefore, the patient population with multiple scans is a highly biased cohort with changing serious illness and it is hard to build predictive models which generalise using this.

1.3.5 Fusion with other data modalities

Fusing CT imaging features from another modality is particularly challenging as CT imaging is extremely high-dimensional, commonly with 50-100 million voxels or more. Therefore, to allow for fusion with much lower dimensional data e.g. clinical variables, feature extraction is typically applied to the CT images by calculation of hand-engineered features or use of a CNN to extract learned features. These lower dimensional radiomic features can be easily combined with the lower dimensional clinical data e.g. by concatenation.

In [68], the authors use a convolutional neural network to extract the features from the CT image and encode the symptoms by convolution. These image features are then combined with the encoded clinical features by multiplication and the

result is concatenated with the image features. This is then passed through several convolutional layers before a prediction is made.

In [40], the authors use a semi-automated algorithm to segment the lung parenchyma into lobes and then threshold to identify the tissue unaffected by ground-glass opacities (GGO) or consolidation. Two radiologists also graded the images for severity. The authors then. build a model using both imaging derived and clinical features, finding that a model using C-reactive protein (an inflammatory marker) along with the imaging features performed better than imaging features alone.

In [66], the authors train a deep learning based segmentation algorithm for GGO and consolidation patterns in CT scans. They fit a model combining radiomic features from these extracted regions with the clinical metadata. This results in a significant boost in performance over using the imaging features alone, increasing the AUC from 0.811 to 0.878.

1.3.6 Common issues with AI and Computed Tomography imaging

Computed Tomography (CT) imaging is a common 3D imaging modality which is typically used for more complex clinical diagnoses due to the high-resolution of the images, a typical voxel in the images is 0.7mm x 0.7mm x 1.0mm. A CT image is reconstructed from 2D X-ray projections and there is no standard algorithm for performing this reconstruction, therefore the images from different scanners can appear quite different qualitatively. Radiomic features extracted from CT images are known to be sensitive to the reconstruction that is used [69].

CT images are commonly enhanced by adding contrast that absorbs X-rays and is added to exaggerate the difference in intensity between adjacent tissues. Therefore, with intensities vastly different between enhanced and non-enhanced images, radiomic features developed for one are unlikely be applicable to the other. Authors must be clear in manuscripts whether enhanced, non-enhanced or a mixture of these are used to train the models (and in what proportions they occur if the latter).

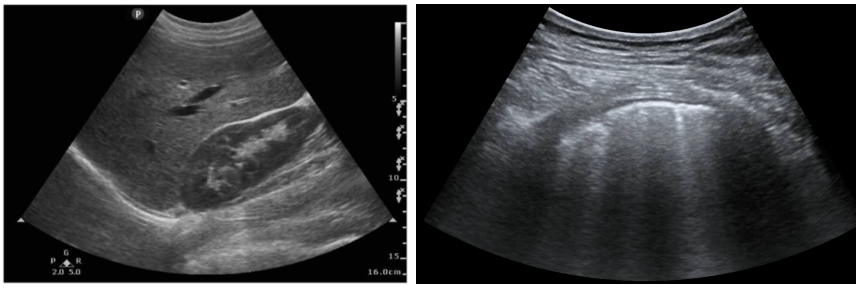
During the COVID-19 pandemic, CT was used in Europe in the more complex cases of COVID-19 where a patient was likely to benefit from it. Therefore, there is an inherent bias in the population of patients who underwent CT screening as they must be ill enough to justify its use (not mild disease) but not so ill that they are ineligible for a CT scan. Also, in the UK and elsewhere, each time a COVID-19 patient used a CT scanner, the machine and the room required a deep clean for 1-2 hours which limited the capacity of hospitals to perform CT scans on COVID-19 patients and hence there is a bias in the data due to this.

CT was used more commonly as a first line imaging modality in China and Russia. Therefore, there are many cases of mild disease within datasets from China and Russia.

1.4 Ultrasound Imaging

The diagnosis and treatment of respiratory diseases rely on the use of various imaging modalities. Chest CT is considered the imaging gold standard for pulmonary diseases, as described in Section 1.3; however, it is expensive and non-portable. Another standard imaging modality utilized to investigate the lung is chest X-ray, which is discussed at length in Section 1.2. Both modalities involve ionizing radiations, which are potentially harmful to the patient. This is particularly significant for specific patient populations such as children, pregnant women, and patients who require repeated examinations over a short period of time. Moreover, CT is generally not available in every hospital nor applicable at bedside, thus requiring patients' mobility. When dealing with a highly infectious disease, this last aspect further increases the risk of contamination within the hospital. Compared to these imaging technologies, ultrasound imaging is safe, cost-effective, more widely available, and transportable, thus has the potential of reaching a much larger population, including non-hospitalized patients. More importantly, there is growing evidence showing that lung ultrasound (LUS) can be used effectively as an imaging modality for pulmonary diseases (e.g., [83, 84, 85, 86, 87]).

However, despite all its advantages LUS has not yet taken a major role in point-of-care protocols. One of the reasons hindering wide-spread use of LUS is the difficulty in interpreting it, which is more challenging compared to other imaging modalities, and even with respect to ultrasonography of other organs. Lungs pose a unique challenge for ultrasound as, normally, ultrasound waves are not transmitted through anatomic structures filled with gas. Consequently, the lung parenchyma is not visible beyond the pleura [83]. This phenomenon is shown in Fig. 1.2: While for other organs (left) ultrasound provides detailed visualizations of the organs, when it comes to the lungs (right) the ultrasound waves are not transmitted through the aerated alveoli and thus no anatomical structure below the pleura is visible in LUS.



(a) US of the abdomen (b) Lung Ultrasound

Fig. 1.2 The challenge of lung Ultrasound (LUS): (a) An ultrasound image of the abdomen, showing the liver and one of the kidneys[a]. (b) A LUS frame showing the pleural line as a bright white curve in the middle of the frame. The entire lung cavity below the pleura is not visible and shows only fog-like noise.

[a] image credit quizlet.com/262841714/liverkidney-interface-diagram

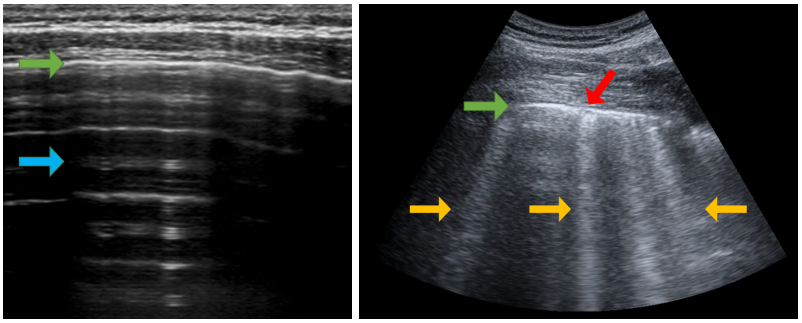
Another key difference between ultrasound and other imaging modalities is the narrow field of view provided by ultrasound. While X-ray and CT can image entire organs and anatomical structures, ultrasound provides only a partial and superficial field of view. As a result, protocols for examining large organs, such as the lungs, require scanning each patient at multiple points, to ensure a complete scan of the organ in question. For LUS protocols vary from, e.g., 6 points [88] to 14 points [86, 89].

What can be observed in LUS

Despite the limitations of LUS, it has been observed that although the visual signal below the pleura fails to show any anatomical structures, it still holds valuable clinical information in the shape of the sonographic artifacts visible in the frame. Fig. 1.3 exemplifies some of these sonographic artifacts.

For example, in normal aerated lung, *A-lines*, hyperechoic, horizontal lines arising at regular intervals from the pleural line can be seen. Fig. 1.3(a) shows an example of A-lines indicated by a blue arrow. A-lines are reverberation artifacts that arise when the ultrasound beam reflects off of the pleura, instead of being absorbed or transmitted through the aerated lung cavity below it. Multiple reverberations result in multiple A-lines, at multiples of the pleural depth. Observing these horizontal A-lines in LUS indicates a well aerated healthy lung.

A different sonographic artifact are vertical *B-lines*, indicated by yellow arrows in Fig. 1.3(b). According to recent developments in LUS [90, 91], vertical artifacts are sonographic signs caused by complex interaction of the multiple scattering phenomena that may form in the presence of an alteration occurring at the lung surface. When forming the LUS frame, the ultrasound signals produced by multiple scattering events, in case of resonance phenomena, are interpreted as a bright vertical



(a) LUS taken using a linear probe. (b) LUS taken using a convex probe.

Fig. 1.3 Examples illustrating what can be observed in LUS using either a linear or convex probe. (a) The scan shows the pleural line (green) and a sequence of horizontal bright A-lines below it (blue). (b) The scan shows the pleural line (green) with some consolidations (red). Several vertical B-lines are visible under the pleural line (yellow). Figure taken from [117].

line emitting from the pleural line and aligned along the ultrasound beam axis [92]. That is, observing B-lines in LUS indicates the presence of fluids just below the pleura. This is, in general, an indication of some pathological condition. The more B-lines observed in a scan or, the wider they are, the more severe is the patient's condition.

Although LUS only shows sonographic artifacts and no anatomical structure below the pleural line, LUS can still be used to inspect the pleural line itself. In healthy patients a smooth and continuous pleural line, as shown in Fig. 1.3(a), is usually observed. In contrast, when there are pleural effusions or pulmonary consolidations the pleural line is observed as irregular and discontinuous, as indicated by a red arrow in Fig. 1.3(b). Based on these observations, it has been sown that LUS can still be very beneficial clinically. For example, [83] showed how LUS can be used for diagnosing the main lung pathologic entities in patients with ARDS, replacing bedside CXR. More recently, the papers [88, 94] demonstrate that LUS may be used to guide COVID-19 patients' management strategies, as well as resource allocation.

To conclude, its widespread availability and safety are putting LUS in a position to take a more substantial role as a POC modality. The main challenge hindering its use is the difficulty in interpreting the acquired frames, inferring the underlying invisible condition indirectly from the visible sonographic artifacts. This usually requires well-trained and highly specialized radiologists. Nevertheless, recent advancements in AI techniques may assist in bridging the gap by processing and analyzing LUS frames, allowing even novice and inexperienced clinicians to benefit from this ubiquitous POC modality.

1.4.1 Models Assisting in Interpreting LUS

Correctly identifying sonographic artifacts, such as A-lines and B-lines, and correctly locating the pleural line is of great importance when one wishes to evaluate a patient's condition from LUS scans. Consequently, several algorithms were developed to detect, segment, and classify these unique LUS features [95, 96, 97, 92], highlighting them to assist clinicians in interpreting LUS scans. Yet, for some of these methods, identifying LUS features is not the end goal but an intermediate stage towards achieving more complicated tasks such as prognosis or diagnosis.

Some algorithms for locating LUS features take advantage of the unique geometric properties of the pleural line, the A- and B- lines and use methods such as Radon-transform [98, 99], dynamic programming [100] or morphological operations [101, 102] to locate them. In contrast, more advanced AI methods rely on training data to accomplish these tasks. These methods range from supervised methods [103, 104, 105, 92], to semi-supervised frameworks [104, 106]. These approaches train deep neural networks for either object detection or semantic segmentation to directly infer the location of the various LUS features. Alternatively, [106] proposes to use gradient-weight class-activation mapping (grad-CAM) [107] on top of a classification network to identify regions of interest in LUS frames.

Ultimately, these approaches automatically highlights the pleural or the A- and B- lines for the clinician during her bed-side examination of the patient, allowing even for a non-expert to benefit from POC LUS modality.

1.4.2 Diagnosis Models

AI can benefit POC LUS beyond just improving human interpretability of the scans. Like in X-ray and CT, AI tools can be used to assist in making differential diagnosis decisions based on LUS. The ultimate goal of these AI tools is assisting in diagnosing a *patient*. This task requires integrating information from several LUS scanning points around the chest, each comprised of many LUS frames. However, most current AI algorithms focus on a less ambitious goal: analyzing only a single LUS frame at a time, leaving the task of integrating the predictions over a sequence of frames and multiple scan points to future works. We focus the discussion in this section and the next one on AI algorithms designed to perform per-frame predictions. In Section 1.4.4 we cover approaches for integrating these per-frame predictions.

To facilitate training of AI models for the task of differential diagnosis of various pulmonary conditions based on LUS, Born *et al.* [108] curated the POCUS dataset. This dataset contains LUS records from mainly 3 different classes, COVID-19, bacterial pneumonia and healthy patients. They gathered LUS scans from different online sources and made it available online[1]. The dataset was constantly updated between April 2020 and January 2021. By January 2021, the dataset contained LUS recordings from 216 patients, where the majority of which acquired with convex transducers and the rest with linear probes. Table 1.1 provides more details about the latest version of the POCUS dataset.

Due to the limited number of LUS scans acquired using a linear probes, and LUS scan of viral pneumonia, the majority of the works that used the POCUS dataset focused only on LUS recordings acquired by convex probes and ignored the non-COVID viral pneumonia [108, 109, 110, 111, 112]. These works focused

Table 1.1 Current POCUS dataset [108]: Number of videos and images per class and probe type.

	Convex		**Linear**	
	Vid.	**Img.**	**Vid.**	**Img.**
COVID-19	64	18	6	4
Bacterial Pneu.	49	20	2	2
Viral Pneu.	3	-	3	-
Healthy	66	15	9	-
Total	**182**	**53**	**20**	**6**

[1] https://github.com/jannisborn/covid19_ultrasound

on developing deep neural networks for LUS frame classification: either classifying each frame into the corresponding diagnosis or providing a binary decision if the frame presents with COVID-19 associated markers or not. Similar to the CXR literature, few papers develop their own custom architectures [112], while most opt to do transfer learning based on existing trained models, from ResNet and Inception backbones to the more efficient MobileNet [108, 109, 108, 113]. Others utilize those models as feature extractors for support vector machine classifier or other classifiers based on fully connected layers [111]. ResNet and Xception architectures reported better performance than the others, with accuracies ranging from 0.83 to 0.99 for frame based diagnosis tasks. However, we caution against direct comparison since the papers not only used different versions of the POCUS dataset but also employed different sampling rate for extracting frames from the dataset's LUS videos.

1.4.3 Prognosis Models

In order to stratify COVID-19 patients using LUS, it has been proposed to score LUS scans according to well-defined physiological findings. Similar scoring scales were independently proposed by [114, 88], suggesting a 4-level scoring system with scores ranging from 0 to 3. Score 0 indicates a healthy lung characterised by a continuous pleural-line and visible A-lines artifacts. In contrast, score 1 indicates first signs of abnormality mostly related to small alterations in the pleural-line, and the appearance of few vertical artifacts. Scores 2 and 3 are representative of a more advanced pathological state, with the presence of small or large consolidations, respectively, and significant presence of vertical artifacts (B-lines and "white lung"). Fig. 1.4 shows example frames representative of each score.

In order to facilitate the development of AI systems to automatically score LUS scans according to their clinical severity, Roy *et al.* [105] curated the Italian COVID-19 Lung Ultrasound dataset (ICLUS)[2]. The ICLUS dataset contains 277 LUS videos of 35 patients, from 5 different Italian medical centers with a total of approximately 60,000 frames acquired by convex and linear probes. All frames in the ICLUS dataset were manually annotated into one of the four severity scores. See Table 1.2 for more details. In addition, video level annotation as well as pixel level annotations for the bio-markers indicative of each score, were provided for a subset of the data.

The ICLUS dataset give rise to several AI methods aiming at scoring LUS frames, thus assisting in the prognosis of the disease. Roy *et al.* [105] utilized the ICLUS dataset for frame-based, video-based and pixel-based severity score prediction. They used a special neural architecture in order to achieve unified processing of LUS frames obtained by either linear or convex probe. In contrast [92] showed that by injecting domain knowledge into the inputs of standard image classification models, following the model-based AI philosophy [93], one can still handle linear and convex frames in a unified manner. They also achieved a boost in performance on two

[2] https://iclus-web.bluetensor.ai/

Table 1.2 ICLUS dataset: Number of LUS frames per severity score and probe type [105]

	Convex	Linear	Total
Score = 0	14,690	5,283	19,973
Score = 1	11,131	3, 164	14,295
Score = 2	15,772	3,200	18,972
Score = 3	3,967	1,717	6,335
Total	**45,560**	**13,364**	**58,924**

domain specific tasks: severity classification and segmentation. Specifically, they injected LUS domain knowledge in the form of B-line and pleural line masks along with the raw LUS frame. Another use of LUS features for prognosis purposes was demonstrated in [100]. Localization of the pleural line was used for extraction of hand-crafted features, such as discontinuities in the pleural line, which in turn were used as an input for a support vector machine classifier of COVID-19 severity score.

A recent dataset comprised over 18,000 LUS frames, acquired by convex probes, from 450 patients (COVID-19 and healthy) was gathered and used for frame based

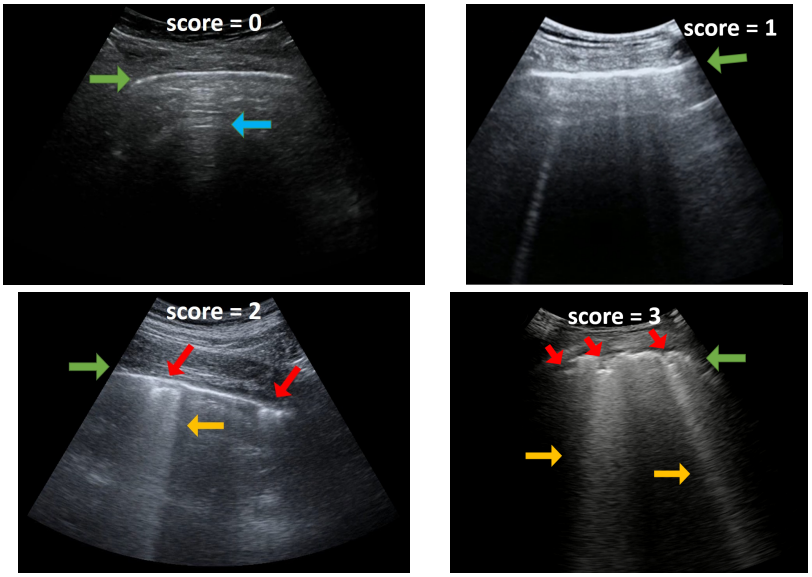


Fig. 1.4 COVID-19 Severity scores. LUS frames exemplifying the severity score of [87] from healthy (score=0, top left) to severe (score=3, bottom right). One can observe the pleural line (green), A-lines (blue), subpleural consolidations (red) and vertical artifacts (e.g., B-lines and "white lung") (yellow). While the pleural line and consolidations are anatomical features, the A-lines and vertical artifacts are sonographic echoes. Figure taken from [92].

.

severity score prediction by [115]. The frames where annotated with the 4 scores severity system proposed in [87] and with a new 7 scores severity system, which gives extra weight to pleural line irregularities. Using Resnet-18 and Resnet-50 they achieved F1 score of above 97% for both models and for both scoring systems. However, this dataset was not made publicly available.

1.4.4 Use of longitudinal imaging

As explained at the beginning of this section, the ultrasound modality in general and LUS, in particular, provides only a partial and limited field of view. Thus, in order to have a complete and comprehensive assessment of the lungs, clinical studies require 6 [88] to 14 scanning points [86, 89]. These multiple scanning points, in turn, amounts to numerous LUS videos, each comprising hundreds of frames. This is in contrast to other modalities in which each scan provides a comprehensive view of the clinical condition. Therefore, when considering longitudinal imaging in the context of LUS, one needs to account for more fundamental levels of information aggregation: The first is accumulating predictions from individual LUS frames to a coherent estimation of the clinical condition at each scanning point. The second level of aggregation is combining the prediction of each scanning point to provide a coherent forecast for the patient as a whole. We discuss these more fundamental levels of longitudinal imaging in LUS next.

The lack of sufficient LUS videos in the available datasets forced most existing AI methods to focus on frame-level predictions. If made, aggregation is done mainly through naïve averaging of the frames' prediction. One exception was introduced by [105], where a lightweight learned approach based on uninorms was used for aggregation of frame-level predictions into video-level prediction showing better performance compared to simple averaging.

The caveat of most current AI methods for LUS, of processing one LUS frame at a time and then aggregating the predictions, is the discarding of temporal information existing in the LUS video sequence. However, exploiting temporal information can be very beneficial in analyzing LUS. Specifically, such information can be helpful in the identification of B-lines due to their flickering nature induced by the motion of the lungs during respiration. For that aim, the use of 3D convolutions, optical flow [116] or long short-term memory (LSTM) layers [110] where the most common approaches.

1.4.5 Common issues with AI and Ultrasound imaging

LUS is acquired using both convex and linear probes. An observation made by [117, 118] suggests that using both in automatic LUS analysis system can be problematic. They noticed that although both anatomic structure and sonographic artifacts can

be observed in both probes, the orientation of the sonographic artifacts changes between probe type. While in a linear probe, B-lines appear "axis-aligned", when using a convex probe, B-lines appear "tilted" as if emitted from the focus point of the probe. This difference has little effect on a human observer, but can be confusing for an automatic LUS analysis system. This observation calls for making an explicit adjustment to the various models, either by new neural architectures (e.g., [105, 117]) or by domain-specific pre-processing of the data (e.g., [92]) in order to make these models suitable across probe types. Alternatively, developing dedicated AI systems restricted to a single probe type hinders the use of other probe types' available training data.

Another issue arising in developing AI methods for LUS is the relatively small amount of available training data. Existing datasets, ICLUS [105] and POCUS [108], have only several thousands of LUS frames. When it comes to LUS videos, or indeed multiple scans of patients, the numbers are significantly lower, in the range of only a few hundreds at best. This quantity does not allow, at the moment, the development of elaborated schemes for information integration over frames in a LUS video or between different videos of the same patient. On the other hand, trying to enrich the datasets by curating LUS data from publicly available sources may also be problematic. LUS frames were available online long before the COVID-19 outbreak, thus acquired by older US machines than the COVID-19 ones. As a result, AI systems trained to perform, e.g., diagnosis, on such data may learn to classify idiosyncrasies related to LUS machines rather than actual clinical findings. This might be the case with some of the results reported on the POCUS dataset.

To conclude this section, LUS has the potential of playing a more significant role as a point-of-care modality. Currently, its interpretability is a challenge that allows only expert clinicians to take advantage of LUS effectively. AI models can potentially bridge the gap and make LUS more widely used. The COVID-19 pandemic has drawn attention to this gap and initiated a surge in the development of relevant AI techniques. These AI models are just starting to emerge; hence they tend to be simplistic, working on single frames rather than aggregating information across frames and scanning points to provide a more holistic prediction for each patient. However, considering the ubiquity of the LUS modality vis-a-vis the difficulty and the expertise needed to gain useful clinical information, it seems like AI can significantly bridge this gap.

1.5 Conclusions

In this section we summarise findings made from the literature, including the success stories of machine learning for the COVID-19 pandemic, the pitfalls which must be given more focus along with the lessons learned and recommendations.

1.5.1 Success Stories

The COVID-19 pandemic highlighted, in many ways, how unprepared the world was to tackle and respond to a highly infectious virus. In particular, for the machine learning and imaging communities, although we had a plethora of literature regarding chest imaging and machine learning, we did not have a robust method to apply to a new disease quickly, that could be validated rapidly. In addition to the modelling issues, data acquisition was also difficult, with lengthy detailed ethics requests required for each hospital to obtain high quality data – along with commonly requiring data extraction by the very same clinicians who are on the front line treating patients. However, we highlight here several of the success stories of this pandemic which serve to provide a blueprint for the next pandemic response.

Large, rapid dataset collection. Most imaging studies outside of COVID-19 are hampered by having small datasets at different sites with images acquired on different machines. This leads to many biases appearing in the development process. The pandemic presented a unique opportunity to allow researchers to collate imaging data at a scale and pace which is commonly not possible with large datasets assembled for many different instruments. The UK's National COVID-19 Chest Imaging Database (NCCID) assembled by NHSX, the British Society of Thoracic Imaging (BSTI), Royal Surrey NHS Foundation Trust and Faculty is a great example of an initiative to collate data in a systematic way which machine learning researchers can use to develop and validate their algorithms. The NCCID required only one umbrella ethics submission for all centers involved and images were collated at a single center for anonymisation and upload into the cloud. Crucially, the burden of the work fell on non-clinical staff to manage the data collation – rightly keeping more clinicians on the front line.

Managing expectations. The pandemic has highlighted areas which were lacking focus in the machine learning for imaging community, namely how to rapidly learn a new class of data from limited examples. It highlighted also how a short-circuit exists in the research community, whereby access to biased public datasets and widely accessible machine learning frameworks (e.g PyTorch, Tensorflow, Keras) allowed a large number of researchers to simultaneously fit models fairly easily – without appreciation for the clinical biases in the data, biases in their methodologies and biases in how models had been evaluated. Through systematic reviews such as [11, 8], the clinical and machine learning community is more aware of the issues pervasive in the machine learning literature, in particular for COVID-19, and can make a fairer assessment of the models being developed, keep their expectations grounded in realty and allow them to ask informed questions.

Appreciation of the need to collaborate. Collaborations across disciplines allow for one community to appreciate the others challenges and understand each others limitations. During the COVID-19 pandemic this has been highlighted the tangible benefits of collaboration. Fundamentally, understanding the clinical pathway for COVID-19 patients is critical to allow for development of high quality models to diagnose and prognosticate for the disease. Understanding when and how ventilators are adjusted, the criteria for patients to be admitted to the intensive care units, when

and where CXR/CT and US imaging are requested and when followup imaging is required are all essential to allow for appreciation of biases in the datasets and methodologies used. Healthy collaborations allow these to be explored in detail and for clinicians to appreciate the limitations of existing machine learning methodologies.

1.5.2 Pitfalls to focus on

In this section we focus on some of the common and continuing areas in which we need to learn and improve for the application of machine learning to images in the time of a pandemic.

Willingness to share data and resources. Crucial to any machine learning task is high-quality data and therefore hospitals and other data contributors need to have a willingness to share data with researchers. For imaging, even if there is willingness, an infrastructure to extract, anonymise and share images is also required. Unfortunately, in the COVID-19 pandemic, even with the NCCID initiative, the hospitals who could share data were those who had capacity for clinicians and informatics staff to extract it. This inclusion bias is critical to tackle as it is necessary that the algorithms which are developed can also apply in those resource stretched hospitals which could not provide data. Federated learning techniques, which do not require sharing of imaging data outside of each institution, such as [71, 119] may be critical in the future to ensure that complications due to sharing and moving of data between sites is not a hindrance to research.

Clearly defining the diagnosis control group. Throughout the literature, diagnosis models for COVID-19 based on machine learning methods rely on distinguishing COVID-19 from a control class or classes. However, it is unclear what this control group should be in the context of COVID-19. Commonly, in the higher quality manuscripts, authors have identified different non-COVID-19 viral pneumonias, bacterial pneumonias and COVID-19 pneumonia as separate classes and trained models to separate these pneumonias. In poorer quality models, healthy patients have been used as controls, or even exclusively paediatric patients with non-COVID-19 pneumonias. The key aspect here is identifying the population in which the model will be applied and ensuring the data used to develop the model reflects this population. For example, if the model is only to be applied by radiologists attempting to distinguish different pneumonias, then training on these alone is reasonable. However, if the model is to be used as a screening tool, the control group should include all different pathologies which are seen in the clinic.

Regulatory considerations. In the literature, many models have been developed and discussed but there is minimal consideration of how these models would pass through regulatory processes to be adopted in the clinic. Many models are described with insufficient documentation to allow for regulation to even be considerable. It is also a reality that regulating software as a medical device is a long and expensive process, which is not only beyond the resources (and skillsets) of many research

groups but means that urgent solutions developed during the pandemic cannot hope to be useful in the course of it due to time delays. It is for policy makers to consider how this particular area of regulation could be improved for the next pandemic.

Missing entries in the data. Missing data is a reality of most real-world clinical datasets and most machine learning models require complete training data. Therefore, it is typical to use data imputation techniques to replace this missing data with intelligent guesses. However, the type of missingness affects the quality of this imputation, i.e. whether the data are missing completely at random, missing at random (with missingness dependent on observed values) or missing not at random [72]. It is important to understand the type of missingness in the data used to develop the methods and ensure that this is also expected in the data to which the model will be applied.

Non-standardised variables across sites. When models are developed for use across multiple sites it is imperative that the variables are comparably recorded at each site (both in terms of the concept being encoded and the units). As machine learning progresses towards more federated approaches with multiple institutions, it is important that variables are mapped into a common standard, e.g. using OMOP dictionaries. Not only does this allow for federated approaches to be adopted more easily, but it ensures that models developed at one site can be rapidly applied at other sites.

1.5.3 Lessons learned and recommendations

The COVID-19 pandemic has highlighted many issues for the community using machine learning with imaging and clinical data. In this section we will highlight several of the lessons learned and recommendations (many of which are covered in [8]).

Recommendations for study design. It is unfortunate that the same degree of care which is applied to clinical trial design is not applied more extensively to machine learning. In particular, exploratory analysis on a small dataset (analagous to a phase 1 study), scaling to include more data with more diversity to determine whether the model is identifying signals in the date (phase 2) before allowing for development on the entirety of the dataset (phase 3). It is most common that authors jump to phase 3, short-cutting a lot of important exploratory steps which can identify biases.

Recommendations for data. Public repositories of imaging for COVID-19 patients should be used with extreme caution. Due to the lack of verification procedures to ensure patients are RT-PCR positive or negative, along with the ability for anyone globally to contribute images, this leads to significant risks of bias (e.g. source issues and Frankenstein datasets) as discussed earlier. Authors must also aim to match demographics across cohorts, an often neglected but significant potential source of bias (which may even be impossible with public datasets that do not include demographic information). Many public datasets obtain their images from preprints and

published manuscripts which are in low-resolution or compressed formats (e.g. JPEG and PNG), rather than their original DICOM format. If this reduction in resolution is biased across the different image classes, this leads to a serious issue for those models reliant on convolutions and hand engineered features which may simply learn to identify the new resolutions.

For CXRs in particular, researchers should be aware that the view (front-to-back vs. back-to-front) that has been used to acquire that CXR is important as, for example, in sick, immobile patients, an front-to-back CXR view is used for practicality rather than the standard back-to-front CXR projection. The most useful algorithms are those that can diagnose disease at an early stage however many datasets will include an overrepresentation of severe disease which will likely reduce the models applicability.

In the literature, the timing between imaging and RT-PCR tests was also largely undocumented, which has implications for the validity of the ground truth used. A negative RT-PCR test does not necessarily mean that a patient does not have COVID-19 and we must encourage authors to evaluate their algorithms on datasets from the pre-COVID-19 era, such as performed by [73], to validate any claims that the algorithm is isolating COVID-19-specific imaging features. In many papers, it is common for non-COVID-19 diagnoses to be determined from the imaging alone, with those same images used to develop the model. This is known as incorporation bias and leads to an over optimistic model performance.

Recommendations for evaluation. The importance of using a well-curated external validation dataset of appropriate size in order to assess generalizability to other cohorts cannot be overstated. Any useful model for diagnosis or prognostication must be robust enough to give reliable results for any sample from the target population rather than just on the sampled population. Calibration statistics should be calculated for the developed models to inform predictive error and decision curve analysis [74] performed for assessing clinical utility. If a model outputs a prediction of death at $p = 0.6$ vs. $p = 0.8$, clinical judgment is likely to change so it is important to know how well calibrated the model is. Authors must also disclose how they ensured that images from the same patient were not included in the different dataset partitions, such as describing patient-level splits. It is primarily an issue for datasets containing multiple images from each patient or those which process 3D volumes as independent 2D samples.

It is important to include confidence intervals, when reporting results, to reflect the uncertainty in the estimate, especially when training models on the small sample sizes commonly seen with COVID-19 data. Moreover, it is important and not an onerous task to demonstrate model interpretability. Examples of interpretability techniques include: (i) informing the clinician of which features in the data most influenced the prediction of the model, (ii) linking the prognostic features to the underlying biology and (iii) overlaying an activation/saliency map on the image to indicate the region of the image which influenced the model's prediction and (iv) identifying patients which had a similar clinical pathway. Many papers derive their performance metrics from the test data alone with an unstated operating point to calculate sensitivity and specificity. Clinical judgment should be used to identify the desired sensitivity or specificity of the model and the operating point should be

derived from the development data. The differences in the sensitivity and specificity of the model should be recorded separately for the validation and test data.

Recommendations for replicability. It is not possible to reproduce many existing models due to updating of publicly available datasets or codes since the publication of the manuscripts. Therefore, we recommend that a cached version of the public dataset be saved, or the date/version quoted, and specific versions of data or code be appropriately referenced. We acknowledge that although perfect replication is potentially not possible, details such as the seeds used for randomness and the actual partitions of the dataset for training, validation and testing would form very useful supplementary materials. Furthermore, it is necessary that the manuscript states any image resizing, cropping and normalisation used to ensure the work is reproducible.

Recommendations for authors. It is recommended that authors assess their manuscript against appropriate established frameworks, such as RQS, CLAIM, TRIPOD, PROBAST and QUADAS [75, 76, 77, 78, 79]. This will ensure reproducibility, and that models are developed in a careful manner.

Recommendations for reviewers. For reviewers, we also recommend the use of the checklists, discussed in the previous point, in order to better identify common weaknesses in reporting the methodology. The most common issues in the papers considered in [8] was the use of biased datasets and/or methodologies. For non-public datasets, it may be difficult for reviewers to assess possible biases if an insufficiently detailed description is given by the authors. We strongly encourage reviewers to ask for clarification from the authors if there is any doubt about bias in the model being considered. Finally, we suggest using reviewers from a combination of both medical and machine learning backgrounds, as they can judge the clinical and technical aspects in different ways.

1.5.4 The next pandemic

For the next pandemic, we must be better prepared, with rapid data collection and sharing, and models which are purpose built and trainable to include new classes quickly and robustly. We need an infrastructure to share data at scale, and we need regulatory improvements which allow for validated algorithms to translate into clinic rapidly.

As a community, we need a blueprint for how to optimally respond to the next pandemic. According to the UK National Audit Office, the COVID-19 pandemic has cost the country over £370 billion as of January 2022 [80] with the testing and contact tracing program alone allocated £37 billion [81]. This astonishing amount highlights the trade-off if governments do not invest in pandemic response research and development to ensure we are not taken by surprise again. A particularly encouraging social enterprise is the Trinity Challenge [82] which ran a competition in 2021 for teams to win funding for their ideas ”to ensure we are better prepared against health emergencies”. There were 8 prize winning teams, all developing tools

using machine learning and data analytics to increase global preparedness for future pandemics.

Acknowledgments

The authors want to thank the whole AIX-COVNET team (see https://covid19ai.maths.cam.ac.uk) and in particular, Sören Dittmer, Julian Gilbey, Matthew Thorpe, Jacobus Preller, Ian Selby, Effrossyni Gkrania-Klotsas, James Rudd, John Aston and Evis Sala, for the many interesting and fruitful discussions and for their many insights on AI methodologies at the interface to clinical practice which have greatly contributed to this article. The authors also thank all their co-authors in papers [27, 92, 117] who worked with them on AI methodologies for COVID-19 detection using Xray and Ultrasound. MR and CBS also acknowledge support from the DRAGON consortium and Intel. MR further acknowledges support from AstraZeneca, and CBS further acknowledges support from her Royal Society Wolfson fellowship, her Philip Leverhulme prize and EPSRC projects EP/N014588/1, EP/T017961/1. SB is a Robin Chemers Neustein Artificial Intelligence Fellow. YE is supported by Miel de Botton, Jean and Terry de Gunzburg Coronavirus Research, a research grant from the Corona Response Fund, and the Manya Igel Centre for Biomedical Engineering and Signal Processing.

References

1. American College of Radiology. (2020, March 22). 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
2. The Royal College of Radiologists. (2020, March 12). The role of CT in patients suspected with COVID-19 infection. https://www.rcr.ac.uk/college/coronavirus-COVID-19-what-rcr-doing/rcr-position-role-ct-patients-suspected-COVID-19
3. Cleverley J., Piper, J., & Jones, M.M.: The role of chest radiography in confirming covid-19 pneumonia. Brit. Med. J., **370**, m2426 (2020)
4. Jacob, J., Alexander, D., Baillie, J.K., Berka, R., Bertolli, O., Blackwood, J., . . . , Joshi, I.: Using imaging to combat a pandemic: Rationale for developing the UK national COVID-19 chest imaging database. Euro. Respir. J., **56**(2). (2020)
5. Harvey, D.M. Example of wallpaper group type p3. Computer-enhanced photograph of a street pavement in Zakopane, Poland. (2005)
6. Driggs, D., Selby, I., Roberts, M., Gkrania-Klotsas, E., Rudd, J. H. F., Yang, G., Babar, J., Sala, E. and Schönlieb, C-B.: Machine Learning for COVID-19 Diagnosis and Prognostication: Lessons for Amplifying the Signal While Reducing the Noise. Radiology: Artificial Intelligence 3(4), e210011 (2021)

7. Rajpurkar, P., Joshi, A., Pareek, A., Chen, P., Kiani, A., Irvin, J., . . . , Lungren, M.P. (2020). CheXpedition: Investigating generalization challenges for translation of chest X-ray algorithms to the clinical setting. Preprint, arXiv:2002.11379.
8. Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., . . . , Schönlieb, C.-B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat. Mach. Int., **3**, 199-217.
9. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C.S., Liang, H., Baxter, S.L., . . . , Zhang, K. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell, 172(5), 1122-1131
10. Inui, S., Fujikawa, A., Jitsu, M., Kunishima, N., Watanabe, S., Suzuki, Y., . . . , Uwabe, Y. (2020). Chest CT findings in cases from the cruise ship Diamond Princess with coronavirus disease (COVID19). Radiol. Cardio. Imag., 2(2).
11. Wynants, L. et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. BMJ 369, m1328 (2020).
12. Hamzeh, A. et al. Artificial intelligence techniques for containment COVID-19 pandemic: a systematic review. Res. Sq. https://doi.org/10.21203/rs.3.rs-30432/v1 (2020).
13. Albahri, O. S. et al. Systematic review of artificial intelligence techniques in the detection and classification of COVID-19 medical images in terms of evaluation and benchmarking: taxonomy analysis, challenges, future solutions and methodological aspects. J. Infect. Public Health https://doi.org/10.1016/j.jiph.2020.06.028 (2020).
14. Feng, S. et al. Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19. IEEE Rev. Biomed. Eng. 14, 4–15 (2021).
15. Gillman, A.G., Lunardo, F., Prinable, J., Belous, G., Nicolson, A., Min, H., Terhorst, A. and Dowling, J.A., 2021. Automated COVID-19 diagnosis and prognosis with medical imaging and who is publishing: a systematic review. Physical and Engineering Sciences in Medicine, pp.1-17.
16. Laino, M.E., Ammirabile, A., Posa, A., Cancian, P., Shalaby, S., Savevski, V. and Neri, E., 2021. The Applications of Artificial Intelligence in Chest Imaging of COVID-19 Patients: A Literature Review. Diagnostics, 11(8), p.1317.
17. Bottino, F., Tagliente, E., Pasquini, L., Napoli, A. D., Lucignani, M., Figà-Talamanca, L., & Napolitano, A. (2021). COVID Mortality Prediction with Machine Learning Methods: A Systematic Review and Critical Appraisal. Journal of personalized medicine, 11(9), 893.
18. Ahuja, V. and Nair, L.V., 2021. Artificial Intelligence and technology in COVID Era: A narrative review. Journal of Anaesthesiology, Clinical Pharmacology, 37(1), p.28.
19. Aishwarya, T. and Kumar, V.R., 2021. Machine Learning and Deep Learning Approaches to Analyze and Detect COVID-19: A Review. SN computer science, 2(3), pp.1-9.
20. Alsharif, M.H., Alsharif, Y.H., Yahya, K., Alomari, O.A., Albreem, M.A. and Jahid, A., 2020. Deep learning applications to combat the dissemination of COVID-19 disease: A review. Eur. Rev. Med. Pharmacol. Sci, 24, pp.11455-11460.
21. Benameur, N., Mahmoudi, R., Zaid, S., Arous, Y., Hmida, B. and Bedoui, M.H., 2021. SARS-CoV-2 diagnosis using medical imaging techniques and artificial intelligence: A review. Clinical Imaging.
22. Bhargava, A. and Bansal, A., 2021. Novel coronavirus (COVID-19) diagnosis using computer vision and artificial intelligence techniques: a review. Multimedia Tools and Applications, pp.1-16.
23. Hariri, W. and Narin, A., 2021. Deep neural networks for COVID-19 detection and diagnosis using images and acoustic-based techniques: a recent review. Soft computing, 25(24), pp.15345-15362.
24. Heidari, A., Navimipour, N.J., Unal, M. and Toumaj, S., 2021. The COVID-19 epidemic analysis and diagnosis using deep learning: A systematic literature review and future directions. Computers in biology and medicine, p.105141.
25. R. J. G. van Sloun, R. Cohen, Y. C. Eldar, "Deep Learning in Ultrasound Imaging", Proceedings of the IEEE, vol. 108, issue 1, pp. 11-29, January 2020.

26. M. Mischi, M. A. Lediju Bell, R. J. G van Sloun and Y. C. Eldar, "Deep Learning in Medical Ultrasound—From Image Formation to Image Analysis", IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 67, issue 12, pp. 2477-2480, December 2020.
27. Keidar, D., Yaron, D., Goldstein, E., Shachar, Y., Blass, A., Charbinsky, L., Aharony, I., Lifshitz, L., Lumelsky, D., Neeman, Z., Mizrachi, M., Hajouj, M., Eizenbach, N., Sela, E., Weiss, C., Levin, P., Benjaminov, O., Bachar, G., Tamir, S., Rapson, Y., Suhami, D., Atar, E., Dror, A., Bogot, N., Grubstein, A., Shabshin, N., Elyada, Y. & Eldar, Y. C. COVID-19 classification of X-ray images using deep neural networks. *European Radiology*. **31**, 9654-9663 (2021)
28. Kaur, J. and Kaur, P., 2021. Outbreak COVID-19 in Medical Image Processing Using Deep Learning: A State-of-the-Art Review. Archives of Computational Methods in Engineering, pp.1-32.
29. Mohammad-Rahimi, H., Nadimi, M., Ghalyanchi-Langeroudi, A., Taheri, M. and Ghafouri-Fard, S., 2021. Application of machine learning in diagnosis of COVID-19 through X-ray and CT images: a scoping review. Frontiers in cardiovascular medicine, 8, p.185.
30. Mondal, M., Bharati, S. and Podder, P., 2021. Diagnosis of COVID-19 Using Machine Learning and Deep Learning: A Review. arXiv preprint arXiv:2110.14910.
31. Montazeri, M., ZahediNasab, R., Farahani, A., Mohseni, H. and Ghasemian, F., 2021. Machine Learning Models for Image-Based Diagnosis and Prognosis of COVID-19: Systematic Review. JMIR medical informatics, 9(4), p.e25181.
32. Ozsahin, I., Sekeroglu, B., Musa, M.S., Mustapha, M.T. and Uzun Ozsahin, D., 2020. Review on diagnosis of COVID-19 from chest CT images using artificial intelligence. Computational and Mathematical Methods in Medicine, 2020.
33. Rezaei, M. and Shahidi, M., 2020. Zero-shot learning and its applications from autonomous vehicles to covid-19 diagnosis: A review. Intelligence-based medicine, p.100005.
34. Soomro, T.A., Zheng, L., Afifi, A.J., Ali, A., Yin, M. and Gao, J., 2021. Artificial intelligence (AI) for medical imaging to combat coronavirus disease (COVID-19): a detailed review with direction for future research. Artificial Intelligence Review, pp.1-31.
35. Salehi, A.W., Baglat, P. and Gupta, G., 2020. Review on machine and deep learning models for the detection and prediction of Coronavirus. Materials Today: Proceedings, 33, pp.3896-3901.
36. A. Sriram*, M. Muckley*, K. Sinha, F. Shamout, J. Pineau, K. J. Geras, L. Azour, Y. Aphinyanaphongs, N. Yakubova, W. Moore. COVID-19 Prognosis via Self-Supervised Representation Learning and Multi-Image Prediction. arXiv preprint arXiv:2101.04909 (2020).
37. Michelle Shu, Richard Strong Bowen, Charles Herrmann, Gengmo Qi, Michele Santacatterina, Ramin Zabih; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4046-4055
38. Kwon, Young Joon, Danielle Toussie, Mark Finkelstein, Mario A. Cedillo, Samuel Z. Maron, Sayan Manna, Nicholas Voutsinas et al. "Combining Initial Radiographs and Clinical Variables Improves Deep Learning Prognostication in Patients with COVID-19 from the Emergency Department." Radiology: Artificial Intelligence 3, no. 2 (2020): e200098.
39. Huang, SC., Pareek, A., Seyyedi, S. et al. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. npj Digit. Med. 3, 136 (2020). https://doi.org/10.1038/s41746-020-00341-z
40. Burian, E.; Jungmann, F.; Kaissis, G.A.; Lohöfer, F.K.; Spinner, C.D.; Lahmer, T.; Treiber, M.; Dommasch, M.; Schneider, G.; Geisler, F.; Huber, W.; Protzer, U.; Schmid, R.M.; Schwaiger, M.; Makowski, M.R.; Braren, R.F. Intensive Care Risk Estimation in COVID-19 Pneumonia Based on Clinical and Imaging Parameters: Experiences from the Munich Cohort. J. Clin. Med. 2020, 9, 1514. https://doi.org/10.3390/jcm9051514
41. Aljouie AF, Almazroa A, Bokhari Y, Alawad M, Mahmoud E, Alawad E, Alsehawi A, Rashid M, Alomair L, Almozaai S, Albesher B, Alomaish H, Daghistani R, Alharbi NK, Alaamery M, Bosaeed M, Alshaalan H. Early Prediction of COVID-19 Ventilation Requirement and Mortality from Routinely Collected Baseline Chest Radiographs, Laboratory, and Clinical Data with Machine Learning. J Multidiscip Healthc. 2021 Jul 30;14:2017-2033. doi: 10.2147/JMDH.S322431. PMID: 34354361; PMCID: PMC8331117.

42. Cohen, J. P., Morrison, P. & Dao, L. COVID-19 image data collection. Preprint at http://arxiv.org/abs/2003.11597 (2020)
43. COVID-19: radiology reference article. Radiopaedia https://radiopaedia.org/articles/covid-19-4?lang=gb (accessed 29 July 2020).
44. COVID-19 Database (SIRM, accessed 29 July 2020); https://www.sirm.org/en/category/articles/covid-19-database/
45. CORONACASES.org (RAIOSS.com, accessed 30 July 2020); https://coronacases.org/
46. Eurorad (ESR, accessed 29 July 2020); https://www.eurorad.org/
47. Chest X-Ray Images (Pneumonia) (Kaggle, accessed 29 July 2020); https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia
48. RSNA Pneumonia Detection Challenge (Kaggle, accessed 29 July 2020); https://www.kaggle.com/c/rsna-pneumonia-detection-challenge
49. Irvin, J. et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. Proc. AAAI Conf. Artif. Intell. 33, 590–597 (2019).
50. Maguolo, G. & Nanni, L. A critic evaluation of methods for COVID-19 automatic detection from X-ray images. Preprint at http://arxiv.org/abs/2004.12823 (2020).
51. M.E.H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M.A. Kadir, Z.B. Mahbub, K.R. Islam, M.S. Khan, A. Iqbal, N. Al-Emadi, M.B.I. Reaz, M. T. Islam, "Can AI help in screening Viral and COVID-19 pneumonia?" IEEE Access, Vol. 8, 2020, pp. 132665 - 132676.
52. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning requires rethinking generalization. In 5th International Conference on Learning Representations ICLR 2017 (ICLR, 2017).
53. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition 248–255 (IEEE, 2010); https://doi.org/10.1109/cvpr.2009.5206848
54. Goodfellow, I. J. et al. Generative adversarial nets. In Proceedings of the 27th International Conference on Neural Information Processing Systems Vol. 2 2672–2680 (MIT Press, 2014).
55. Acar, E., Şahin, E. & Yilmaz, İ. Improving effectiveness of different deep learning-based models for detecting COVID-19 from computed tomography (CT) images. Preprint at medRxiv https://doi.org/10.1101/2020.06.12.20129643 (2020).
56. Bae S, Kim SR, Kim M, et alImpact of cardiovascular disease and risk factors on fatal outcomes in patients with COVID-19 according to age: a systematic review and meta-analysisHeart 2021;107:373-380.
57. Grodecki K, Lin A, Razipour A, Cadet S, McElhinney PA, Chan C, Pressman BD, Julien P, Maurovich-Horvat P, Gaibazzi N, Thakur U, Mancini E, Agalbato C, Menè R, Parati G, Cernigliaro F, Nerlekar N, Torlasco C, Pontone G, Slomka PJ, Dey D. Epicardial adipose tissue is associated with extent of pneumonia and adverse outcomes in patients with COVID-19. Metabolism. 2021 Feb;115:154436. doi: 10.1016/j.metabol.2020.154436. Epub 2020 Nov 19. PMID: 33221381; PMCID: PMC7676319.
58. Stancil, I.T., Michalski, J.E., Davis-Hall, D. et al. Pulmonary fibrosis distal airway epithelia are dynamically and structurally dysfunctional. Nat Commun 12, 4566 (2021). https://doi.org/10.1038/s41467-021-24853-8
59. Michael Roberts, Kirl Kirov, Tom Mclellan, Evan Morgan, Fahdi Kanavati, Darren Gallagher, Philip Molyneaux, Carola-Bibane Schönlieb, Alessandro Ruggiero, Muhunthan Thillai. Late Breaking Abstract - Fully automated airway measurement correlates with radiological disease progression in Idiopathic Pulmonary Fibrosis. European Respiratory Journal Sep 2021, 58 (suppl 65) OA3951; DOI: 10.1183/13993003.congress-2021.OA3951
60. Medeiros, A. K., Barbisan, C. C., Cruz, I. R., de Araújo, E. M., Libânio, B. B., Albuquerque, K. S., & Torres, U. S. (2020). Higher frequency of hepatic steatosis at CT among COVID-19-positive patients. Abdominal radiology (New York), 45(9), 2748–2754. https://doi.org/10.1007/s00261-020-02648-7
61. Chen, L., Wang, Q., Wu, H., Hu, J., & Zhang, J. (2020). REPEAT CHEST CT SCANS IN MODERATE-TO-SEVERE PATIENTS' MANAGEMENT DURING THE COVID-19 PANDEMIC: OBSERVATIONS FROM A SINGLE CENTRE IN WUHAN, CHINA. Radiation protection dosimetry, 190(3), 269–275. https://doi.org/10.1093/rpd/ncaa106

62. Jevnikar, M., Sanchez, O., Chocron, R., Andronikof, M., Raphael, M., Meyrignac, O., Fournier, L., Montani, D., Planquette, B., Soudani, M. and Boucly, A., 2021. Prevalence of pulmonary embolism in patients with COVID 19 at the time of hospital admission. European Respiratory Journal.
63. Sonia Baeza, Roger Domingo, Maite Salcedo-Pujantell, Jordi Deportós, Gloria Moragas, Ignasi Garcia-Olivé, Carles Sanchez, Debora Gil, Antoni Rosell. Artificial intelligence to optimize pulmonary embolism diagnosis during covid-19 pandemic by perfusion SPECT/CT, a pilot study. European Respiratory Journal Sep 2021, 58 (suppl 65) PA359; DOI: 10.1183/13993003.congress-2021.PA359
64. https://www.bsti.org.uk/media/resources/files/BSTI_COVID-19_Radiology_Guidance_version_2_16.03.20.pdf
65. Gao, Yang MD, PhD; Hu, Yuxiong MD; Zhu, Junteng MD; Liu, Huan MD; Qiu, Rongxian MD; Lin, Qunying MD; He, Xiongzhi MD; Lin, Hai-Bin MD; Cheng, Shimingg; Li, Guangxi MD. The value of repeated CT in monitoring the disease progression in moderate COVID-19 pneumonia, Medicine: March 12, 2021 - Volume 100 - Issue 10 - p e25005 doi: 10.1097/MD.0000000000025005
66. Zhang, K., Liu, X., Shen, J., Li, Z., Sang, Y., Wu, X., Zha, Y., Liang, W., Wang, C., Wang, K. and Ye, L., 2020. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. Cell, 181(6), pp.1423-1433.
67. Wang, G., Liu, X., Shen, J., Wang, C., Li, Z., Ye, L., Wu, X., Chen, T., Wang, K., Zhang, X. and Zhou, Z., 2021. A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images. Nature Biomedical Engineering, 5(6), pp.509-521.
68. Huang, Z., Liu, X., Wang, R., Zhang, M., Zeng, X., Liu, J., Yang, Y., Liu, X., Zheng, H., Liang, D. and Hu, Z., 2021. FaNet: fast assessment network for the novel coronavirus (COVID-19) pneumonia based on 3D CT imaging and clinical symptoms. Applied Intelligence, 51(5), pp.2838-2849.
69. Orlhac, F., Frouin, F., Nioche, C., Ayache, N. and Buvat, I., 2019. Validation of a method to compensate multicenter effects affecting CT radiomics. Radiology, 291(1), pp.53-59.
70. Scott, Lauren J., et al. "Association between National Early Warning Scores in primary care and clinical outcomes: an observational study in UK primary and secondary care." British Journal of General Practice 70.695 (2020): e374-e380.
71. Bai, X., Wang, H., Ma, L. et al. Advancing COVID-19 diagnosis with privacy-preserving collaboration in artificial intelligence. Nat Mach Intell 3, 1081–1089 (2021). https://doi.org/10.1038/s42256-021-00421-z
72. White, Ian R., Patrick Royston, and Angela M. Wood. "Multiple imputation using chained equations: issues and guidance for practice." Statistics in medicine 30.4 (2011): 377-399.
73. Banerjee, Imon, et al. "Was there COVID-19 back in 2012? Challenge for AI in diagnosis with similar indications." arXiv preprint arXiv:2006.13262 (2020).
74. Vickers, Andrew J., and Elena B. Elkin. "Decision curve analysis: a novel method for evaluating prediction models." Medical Decision Making 26.6 (2006): 565-574.
75. Lambin, P. et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat. Rev. Clin. Oncol. 14, 749–762 (2017).
76. Mongan, J., Moy, L. & Kahn, C. E. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol. Artif. Intell. 2, e200029 (2020).
77. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann. Intern. Med. 162, 55–63 (2015).
78. Wolff, R. F. et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann. Intern. Med. 170, 51 (2019).
79. Whiting, P. F. et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann. Intern. Med. 155, 529 (2011).
80. https://www.nao.org.uk/covid-19/cost-tracker/ (accessed 31st Jan 2022).

81. https://committees.parliament.uk/publications/4976/documents/50058/default/ (accessed 31st Jan 2022).
82. https://thetrinitychallenge.org/ (accessed 31st Jan 2022).
83. Lichtenstein, D., Goldstein, I., Mourgeon, E., Cluzel, P., Grenier, P. & Rouby, J. Comparative diagnostic performances of auscultation, chest radiography, and lung ultrasonography in acute respiratory distress syndrome. *The Journal Of The American Society Of Anesthesiologists*. (2004)
84. Allinovi, M., Parise, A., Giacalone, M., Amerio, A., Delsante, M., Odone, A., Franci, A., Gigliotti, F., Amadasi, S., Delmonte, D. & Others Lung ultrasound may support diagnosis and monitoring of COVID-19 pneumonia. *Ultrasound In Medicine And Biology*. **46**, 2908-2917 (2020)
85. Kameda, T., Mizuma, Y., Taniguchi, H., Fujita, M. & Taniguchi, N. Point-of-care lung ultrasound for the assessment of pneumonia: a narrative review in the COVID-19 era. *Journal Of Medical Ultrasonics*. pp. 1-13 (2021)
86. Smargiassi, A., Soldati, G., Torri, E., Mento, F., Milardi, D., Giacomo, P., De Matteis, G., Burzo, M., Larici, A., Pompili, M. & Others Lung ultrasound for COVID-19 patchy pneumonia: extended or limited evaluations?. *Journal Of Ultrasound In Medicine*. (2020)
87. Soldati, G., Smargiassi, A., Inchingolo, R., Buonsenso, D., Perrone, T., Briganti, D., Perlini, S., Torri, E., Mariani, A., Mossolani, E. & Others Proposal for International Standardization of the Use of Lung Ultrasound for Patients With COVID-19: A Simple, Quantitative, Reproducible Methodf. *Journal Of Ultrasound In Medicine*. (2020)
88. Lichter, Y., Topilsky, Y., Taieb, P., Banai, A., Hochstadt, A., Merdler, I., Oz, A., Vine, J., Goren, O., Cohen, B. & Others Lung ultrasound predicts clinical course and outcomes in COVID-19 patients. *Intensive Care Medicine*. (2020)
89. Mento, F., Perrone, T., Macioce, V., Tursi, F., Buonsenso, D., Torri, E., Smargiassi, A., Inchingolo, R., Soldati, G. & Demi, L. On the Impact of different lung ultrasound imaging protocols in the evaluation of patients affected by coronavirus disease 2019: How many acquisitions are needed?. *Journal Of Ultrasound In Medicine*. (2020)
90. Mento, F. & Demi, L. On the influence of imaging parameters on lung ultrasound B-line artifacts, in vitro study. *The Journal Of The Acoustical Society Of America*. (2020)
91. Demi, L., Demi, M., Prediletto, R. & Soldati, G. Real-time multi-frequency ultrasound imaging for quantitative lung ultrasound – first clinical results. *The Journal Of The Acoustical Society Of America*. (2020)
92. Frank, O., Schipper, N., Vaturi, M., Soldati, G., Smargiassi, A., Inchingolo, R., Torri, E., Perrone, T., Mento, F., Demi, L., Galun, M., Eldar, Y. C. & Bagon, S. Integrating Domain Knowledge into Deep Networks for Lung Ultrasound with Applications to COVID-19. *IEEE Transactions On Medical Imaging*. (2021)
93. Shlezinger, N., Whang, J., Eldar, Y. C. & Dimakis, A. G. Model-Based Deep Learning. *Submitted to IEEE Transactions on Signal Processing*. (2022)
94. Perrone, T., Soldati, G., Padovini, L., Fiengo, A., Lettieri, G., Sabatini, U., Gori, G., Lepore, F., Garolfi, M., Palumbo, I. & Others A new lung ultrasound protocol able to predict worsening in patients affected by severe acute respiratory syndrome coronavirus 2 pneumonia. *Journal Of Ultrasound In Medicine*. (2020)
95. McDermott, C., Lacki, M., Sainsbury, B., Henry, J., Filippov, M. & Rossa, C. Sonographic diagnosis of COVID-19: A review of image processing for lung ultrasound. *Frontiers In Big Data*. **4** pp. 2 (2021)
96. Mento, F., Soldati, G., Prediletto, R., Demi, M. & Demi, L. Quantitative lung ultrasound spectroscopy applied to the diagnosis of pulmonary fibrosis: The first clinical study. *IEEE Transactions On Ultrasonics, Ferroelectrics, And Frequency Control*. **67**, 2265-2273 (2020)
97. Wang, Y., Zhang, Y., He, Q., Liao, H. & Luo, J. A semi-automatic ultrasound image analysis system for the grading diagnosis of COVID-19 pneumonia. *ArXiv Preprint ArXiv:2111.02676*. (2021)
98. Anantrasirichai, N., Hayes, W., Allinovi, M., Bull, D. & Achim, A. Line detection as an inverse problem: application to lung ultrasound imaging. *IEEE Transactions On Medical Imaging*. (2017)

99. Karakuş, O., Anantrasirichai, N., Aguersif, A., Silva, S., Basarab, A. & Achim, A. Detection of Line Artifacts in Lung Ultrasound Images of COVID-19 Patients Via Nonconvex Regularization. *IEEE Transactions On Ultrasonics, Ferroelectrics, And Frequency Control*. (2020)
100. Carrer, L., Donini, E., Marinelli, D., Zanetti, M., Mento, F., Torri, E., Smargiassi, A., Inchingolo, R., Soldati, G., Demi, L. & Others Automatic pleural line extraction and COVID-19 scoring from lung ultrasound data. *IEEE Transactions On Ultrasonics, Ferroelectrics, And Frequency Control*. (2020)
101. Moshavegh, R., Hansen, K., Sørensen, H., Hemmsen, M., Ewertsen, C., Nielsen, M. & Jensen, J. Novel automatic detection of pleura and B-lines (comet-tail artifacts) on in vivo lung ultrasound scans. *Medical Imaging 2016: Ultrasonic Imaging And Tomography*. **9790** pp. 97900K (2016)
102. Brusasco, C., Santori, G., Bruzzo, E., Trò, R., Robba, C., Tavazzi, G., Guarracino, F., Forfori, F., Boccacci, P. & Corradi, F. Quantitative lung ultrasonography: a putative new algorithm for automatic detection and quantification of B-lines. *Critical Care*. **23**, 1-7 (2019)
103. Kulhare, S., Zheng, X., Mehanian, C., Gregory, C., Zhu, M., Gregory, K., Xie, H., Jones, J. & Wilson, B. Ultrasound-based detection of lung abnormalities using single shot detection convolutional neural networks. *Simulation, Image Processing, And Ultrasound Systems For Assisted Diagnosis And Navigation*. pp. 65-73 (2018)
104. Mason, H., Cristoni, L., Walden, A., Lazzari, R., Pulimood, T., Grandjean, L., Wheeler-Kingshott, C., Hu, Y. & Baum, Z. Lung Ultrasound Segmentation and Adaptation Between COVID-19 and Community-Acquired Pneumonia. *International Workshop On Advances In Simplifying Medical Ultrasound*. pp. 45-53 (2021)
105. Roy, S., Menapace, W., Oei, S., Luijten, B., Fini, E., Saltori, C., Huijben, I., Chennakeshava, N., Mento, F., Sentelli, A. & Others Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound. *IEEE Transactions On Medical Imaging*. (2020)
106. Sloun, R. & Demi, L. Localizing B-lines in lung ultrasonography by weakly supervised deep learning, in-vivo results. *IEEE Journal Of Biomedical And Health Informatics*. **24**, 957-964 (2019)
107. Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings Of The IEEE International Conference On Computer Vision*. (2017)
108. Born, J., Wiedemann, N., Cossio, M., Buhre, C., Brändle, G., Leidermann, K., Aujayeb, A., Moor, M., Rieck, B. & Borgwardt, K. Accelerating Detection of Lung Pathologies with Explainable Ultrasound Image Analysis. *Applied Sciences*. (2021)
109. Awasthi, N., Dayal, A., Cenkeramaddi, L. & Yalavarthy, P. Mini-COVIDNet: Efficient Lightweight Deep Neural Network for Ultrasound Based Point-of-Care Detection of COVID-19. *IEEE Transactions On Ultrasonics, Ferroelectrics, And Frequency Control*. **68**, 2023-2037 (2021)
110. Barros, B., Lacerda, P., Albuquerque, C. & Conci, A. Pulmonary covid-19: Learning spatiotemporal features combining cnn and lstm networks for lung ultrasound video classification. *Sensors*. **21**, 5486 (2021)
111. Al-Jumaili, S., Duru, A. & Uçan, O. Covid-19 Ultrasound image classification using SVM based on kernels deduced from Convolutional neural network. *2021 5th International Symposium On Multidisciplinary Studies And Innovative Technologies (ISMSIT)*. pp. 429-433 (2021)
112. Muhammad, G. & Hossain, M. COVID-19 and non-COVID-19 classification using multi-layers fusion from lung ultrasound images. *Information Fusion*. **72** pp. 80-88 (2021)
113. Diaz-Escobar, J., Ordóñez-Guillén, N., Villarreal-Reyes, S., Galaviz-Mosqueda, A., Kober, V., Rivera-Rodriguez, R. & Lozano Rizk, J. Deep-learning based detection of COVID-19 using lung ultrasound imagery. *Plos One*. **16**, e0255886 (2021)
114. Soldati, G., Smargiassi, A., Inchingolo, R., Buonsenso, D., Perrone, T., Briganti, D., Perlini, S., Torri, E., Mariani, A., Mossolani, E. & Others Is there a role for lung ultrasound during the COVID-19 pandemic?. *Journal Of Ultrasound In Medicine*. (2020)

115. La Salvia, M., Secco, G., Torti, E., Florimbi, G., Guido, L., Lago, P., Salinaro, F., Perlini, S. & Leporati, F. Deep learning and lung ultrasound for Covid-19 pneumonia detection and severity classification. *Computers In Biology And Medicine*. **136** pp. 104742 (2021)
116. Krishnaswamy, D., Ebadi, S., Bolouri, S., Zonoobi, D., Greiner, R., Meuser-Herr, N., Jaremko, J., Kapur, J., Noga, M., Punithakumar, K. & Others A novel machine learning-based video classification approach to detect pneumonia in COVID-19 patients using lung ultrasound. *International Journal Of Noncommunicable Diseases*. **6**, 69 (2021)
117. Yaron, D., Keidar, D., Goldstein, E., Shachar, Y., Blass, A., Frank, O., Schipper, N., Shabshin, N., Grubstein, A., Suhami, D., Bogot, N. R., Weiss, C. S., Sela, E., Dror, A. A., Vaturi, M., Mento, F., Torri, E., Inchingolo, R., Smargiassi, A., Soldati, G., Perrone, T., Demi, L., Galun, M., Bagon, S., Elyada, Y. M. & Eldar, Y. C. Point of Care Image Analysis for COVID-19. *IEEE International Conference On Acoustics, Speech And Signal Processing*. (2021)
118. Bagon, S., Galun, M., Frank, O., Schipper, N., Vaturi, M., Zalcberg, G., Soldati, G., Smargiassi, A., Inchingolo, R., Torri, E., Perrone, T., Mento, F., Demi, L. & Eldar, Y. C. Assessment of COVID-19 in lung ultrasound by combining anatomy and sonographic artifacts using deep learning. *The Journal Of The Acoustical Society Of America*. (2020), tinyurl.com/yxcwaer1
119. T. Gafni, N. Shlezinger, K. Cohen, Y. C. Eldar & H. V. Poor Federated Learning: A Signal Processing Perspective. *to appear in Signal Processing Magazine*. (2022)