Expert Systems With Applications 235 (2024) 121089

Contents lists available at ScienceDirect

Expert Systems With Applications

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

# A deep learning method for predicting the COVID-19 ICU patient outcome fusing X-rays, respiratory sounds, and ICU parameters

Yunan Wu [a,*], Bruno Machado Rocha [b], Evangelos Kaimakamis [c], Grigorios-Aris Cheimariotis [d], Georgios Petmezas [d], Evangelos Chatzis [d], Vassilis Kilintzis [d], Leandros Stefanopoulos [d], Diogo Pessoa [b], Alda Marques [e], Paulo Carvalho [b], Rui Pedro Paiva [b], Serafeim Kotoulas [c], Militsa Bitzani [c], Aggelos K. Katsaggelos [a], Nicos Maglaveras [d]

[a] The Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL, USA
[b] The Department of Informatics Engineering, Centre for Informatics and Systems of the University of Coimbra, The University of Coimbra, Coimbra, Portugal
[c] 1st Intensive Care Unit, "G. Papanikolaou" General Hospital of Thessaloniki, Greece
[d] 2nd Department of Obstetrics and Gynaecology, Laboratory of Computing, Medical Informatics and Biomedical - Imaging Technologies, Medical School, Aristotle University of Thessaloniki, Greece
[e] The Lab3R — Respiratory Research and Rehabilitation Laboratory, School of Health Sciences (ESSUA), and the Institute of Biomedicine (iBiMED), University of Aveiro, Aveiro, Portugal

## ARTICLE INFO

Dataset link: https://figshare.com/s/e5af036d5ca46150eac4

*Keywords:*
COVID-19
Deep learning fusion
Respiratory sounds
Clinical variables
Chest X-rays
ICU mortality

## ABSTRACT

Assessing the health status of critically ill patients with COVID-19 and predicting their outcome are highly challenging problems and one of the reasons for poor management of ICU resources worldwide. A better pathophysiological understanding of patients' state evolution in the ICU can enhance effective medical interventions. Therefore, there is a need to monitor and analyze the pulmonary function of a ICU patient with COVID-19 and its impact on cardiovascular and other systems. To achieve this, chest X-rays (CXRs), respiratory sounds and all the routinely monitored parameters, scores and metrics in the COVID-19 ICU were recorded from 171 ICU patients with COVID-19 from June 2020 until December 2021. Features were extracted from respiratory sounds, deep learning analysis was conducted on CXRs, and logistic regression analysis was performed on routine ICU clinical variables. Deep learning pipelines were established to classify patients' outcomes (survival or death) at two time points (ICU mortality or 90-day mortality) using three input configurations: (a) CXRs, (b) a fusion of CXRs and respiratory sounds features, or (c) a fusion of CXRs, respiratory sounds features, and principal features of the ICU clinical measurements. The performance of the latter approach was promising, achieving, for ICU mortality, an accuracy of 0.761 and an AUC of 0.759, and for 90-day mortality, an accuracy of 0.743 and an AUC of 0.752, while the performance of approaches (a) and (b) was worse. Therefore, using multi-source data and longitudinal COVID-19 ICU data offers a better prediction of the outcome in the ICU, thereby optimizing medical decisions and interventions. Furthermore, we show that adding the adventitious respiratory sounds features significantly increased AUC and accuracy for mortality prediction of ICU patients with COVID-19.

## 1. Introduction

The COVID-19 pandemic has caused millions of deaths worldwide and overloaded the healthcare systems around the world (Velavan & Meyer, 2020). During the pandemic, and currently intensive care units (ICUs) have treated millions of patients with COVID-19, with a large percentage dying after irreversible damage in the lungs and other vital body functions and structures (Armstrong, Kane, Kursumovic, Oglesby, & Cook, 2021). To a large extent, the evolution of life

* Correspondence to: 2145 Sheridan Road, Evanston, IL, 60201, USA.
*E-mail addresses:* yunanwu2020@u.northwestern.edu (Y. Wu), bmrocha@dei.uc.pt (B.M. Rocha), vakaimak@yahoo.gr (E. Kaimakamis), ncheimar@gmail.com (G.-A. Cheimariotis), petmezgs@auth.gr (G. Petmezas), chatzise@auth.gr (E. Chatzis), billyk@auth.gr (V. Kilintzis), lstefano@auth.gr (L. Stefanopoulos), dpessoa@dei.uc.pt (D. Pessoa), amarques@ua.pt (A. Marques), carvalho@dei.uc.pt (P. Carvalho), ruipedro@dei.uc.pt (R.P. Paiva), akiskotoulas@hotmail.com (S. Kotoulas), bitmilly@gmail.com (M. Bitzani), a-katsaggelos@northwestern.edu (A.K. Katsaggelos), nicmag@auth.gr (N. Maglaveras).

https://doi.org/10.1016/j.eswa.2023.121089
Received 5 March 2023; Received in revised form 27 June 2023; Accepted 29 July 2023
Available online 7 August 2023

threatening Acute Respiratory Distress Syndrome (ARDS) caused by the associated coronavirus (SARS-CoV-2) is not well understood as concerns its pathophysiology (Yuki, Fujiogi, & Koutsogiannaki, 2020). Efforts have been made to predict the final outcome of the ICU stay of critically ill patients with COVID-19 based on clinical and laboratory findings with variable results (Serafim, Póvoa, Souza-Dantas, Kalil, & Salluh, 2021). These efforts often failed to take into account the complex correlations between certain clinical manifestations of severe ARDS and the clinical course of the disease in the controlled ICU environment. A predictive algorithm capable of discriminating between patients at higher risk of death based on early clinical respiratory functional parameters and findings from simple examinations using available medical devices (like pulmonary auscultation and chest X-rays) would likely lead to more effective management of high-risk patients and potentially to an increase in overall survival. Ideally, this algorithm should be easy to feed, include meaningful bio-parameters or imaging modalities, and have adequate precision rates.

Over the past two years, deep learning (DL) has played an important role in the detection of patients with COVID-19. Trained on large amounts of computed tomography (CT) scans or chest X-rays (CXRs), DL models are able to diagnose COVID-19 faster and more accurately than radiologists (Tabik et al., 2020; Wehbe et al., 2021). For example, Ramsey et al. developed a DeepCOVID-XR algorithm that integrates six different DL models to detect COVID-19 on CXRs, which outperformed experienced radiologists (Wehbe et al., 2021). However, so far, only a few studies have made an attempt to predict the clinical course of patients with COVID-19 using DL techniques. In particular, Sriram et al. (2021) used CXRs to propose a self-supervised method based on the DenseNet-121 architecture (Huang, Liu, Van Der Maaten, & Weinberger, 2017) for the prediction of COVID-19 patient deterioration including three different tasks: namely, adverse events (AUC = 0.742) prediction from a single chest radiograph, increased oxygen requirements (AUC = 0.765) prediction from a single chest radiograph, and adverse events (AUC = 0.786) and mortality (AUC = 0.848) prediction from a sequence of radiographs. On the other hand, Shamout et al. (2021) presented a DL approach that combines a deep CNN for CXRs feature extraction and a gradient boosting model for routine clinical parameter learning to predict the deterioration risk of patients with COVID-19 (AUC = 0.786). Similarly, Kwon et al. (2021) fused chest radiographs and clinical variables into a DenseNet-121 architecture to predict the intubation (AUC = 0.88) and mortality (AUC = 0.82) risk of patients with COVID-19, while Aljouie et al. (2021) tested four different machine learning classifiers for the prediction of ventilation requirement (AUC = 0.87) and mortality risk (AUC = 0.83) using a fusion of CXRs, complete blood count, demographic and clinical data. Finally, Gourdeau et al. (2022) applied transfer learning to extract meaningful features from COVID-19 CXRs and predict the outcome of mechanical ventilation and achieved an AUC of 0.702 using only pre-intubation CXRs and an AUC of 0.743 when combining imaging data and aggregated risk factors. However, with the exception of the last publication (Gourdeau et al., 2022), all researchers have tried to predict the clinical outcome of non-severely ill patients with COVID-19 in settings outside the ICU. The inherently complex nature of critically ill patients with COVID-19 related ARDS requires fusion of clinical physiology parameters together with imaging and other clinical examination data to allow for a more robust and reliable prognostic sequence.

Apart from researchers who used DL techniques to predict mortality in severely ill patients with COVID-19, various authors have described clinical models capable of predicting the in-hospital mortality of patients with COVID-19 in the ICU. Most of these efforts have identified the following clinical parameters that are associated with increased probability of death in the ICU: increased age (Alser et al., 2021; Ferrando et al., 2020; Gallo Marin et al., 2021), presence of severe ARDS (Alser et al., 2021; Ferrando et al., 2020), high sequential organ failure assessment (SOFA) score (Alser et al., 2021; Ferrando et al., 2020; Gallo Marin et al., 2021), extensive lung involvement in CT scans, specific biomarkers (Gallo Marin et al., 2021) and complications during the ICU stay (like acute kidney injury, septic shock, cardiac arrhythmias, and infections) (Ferrando et al., 2020). Another study found that 18 day-mortality was associated with increased age, obesity, high SOFA score and low ratio of partial arterial oxygen pressure and inspiratory oxygen fraction (PaO2/FiO2 ratio) (Leoni et al., 2021). These efforts have highlighted the importance of specific clinical parameters that play an important role in the mortality prediction for this pool of patients.

The objective of this work was to make mortality predictions for ICU patients with COVID-19 by integrating CXRs, clinical variables, and respiratory sounds features. The contribution of this paper can be summarized as follows: (1) This study represents the first attempt, to the best of our knowledge, to merge these three modalities to predict mortality in the ICU setting for COVID-19 patients. By combining information from CXRs, clinical variables, and respiratory sounds, we aimed to enhance the accuracy and reliability of mortality predictions. The analysis of sounds from a clinical point of view can enable physicians to determine the progression of the lung insult and in some cases, the severity of the lung abnormalities, especially in the presence of certain adventitious sounds, like crackles and squawks. However, the doctors cannot tell the difference between the patients who will survive and those who will not base on the respiratory sounds alone despite the important information they provide, thus an automated algorithm taking into account the respiratory sounds, chest X-rays and clinical parameters would be ideal for a more reliable prediction of the ICU outcome of COVI-19 patients; (2) This study introduces a benchmark deep learning pipeline for the outcome prediction of COVID-19 ICU patients; (3) Importantly, we established the first multimodal (CXRs + respiratory sound + ICU parameters) open access database of ICU COVID-19 patients, which will facilitate further research and promote the development of new methodologies in this domain. Therefore, the paper is structured as follows: in Section 2, the database, the preprocessing steps, and the proposed method are first presented; in Section 3, the obtained results are analyzed and lastly, in Section 4, the strengths and limitations of the study are discussed.

## 2. Materials and methods

In this study, we developed a multi-model fusion workflow to predict mortality for ICU patients by integrating CXRs, clinical variables, and respiratory sound features, as shown in Fig. 1. The data from these three modalities were collected at sequential time points and underwent a pre-processing stage to ensure optimal quality. Specifically, the fusion model was initialized using pre-trained weights derived from a variational auto-encoder (VAE), providing a solid starting point for the fusion process. To capture temporal patterns and extract valuable insights, we employed long short-term memory (LSTM) networks within the fusion model, enabling the extraction of temporal features over time.

### 2.1. Dataset and ethics

The dataset was collected by an ICT platform that enables the monitoring and fusion of clinical information from patients with COVID-19 admitted to the ICU into an annotated database named CoCross (Kilintzis et al., 2022). The CoCross platform was deployed in June 2020 in the 1st ICU of "G. Papanikolaou" hospital in Thessaloniki, Greece, where recordings from patients with COVID-19 receiving care in ICU were performed. The study protocol was approved by the Ethics Committee of the hospital (Scientific Council of "G. Papanikolaou" Hospital, Approval Number: 42/10/20-05-2020). Due to the absence of additional interventions during the study and the special circumstances of the COVID-19 pandemic (prohibition of relatives' visit to the ICU), the Ethics Committee waived the need for written consent form for the

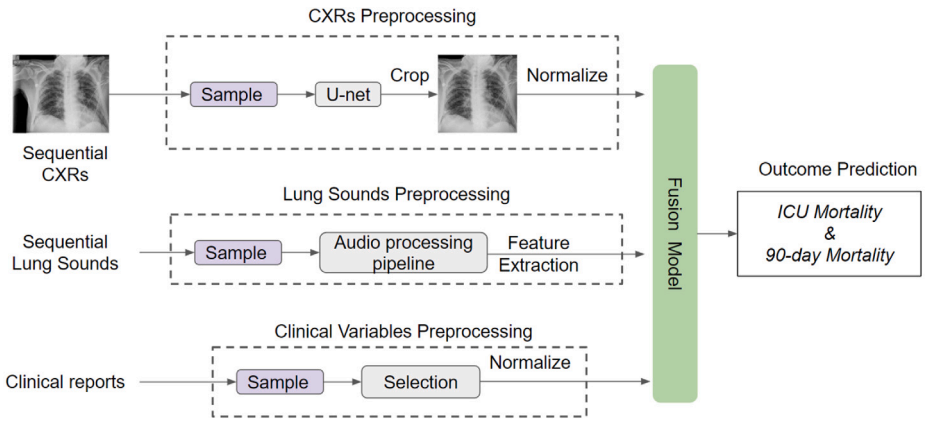


**Fig. 1.** An overview of the entire multi-modal fusion workflow.

recordings, provided that all the current regulations regarding data protections are followed. In total, multi-source data from 171 ICU patients (female: male = 57: 114) have been acquired and imported in the CoCross database using FHIR based data modeling (Kilintzis, Chouvarda, Beredimas, Natsiavas, & Maglaveras, 2019), corresponding to a five-day average monitoring period including a dataset with 3477 distinct auscultations. Specifically, in each ICU segment a 10″ tablet, a Bluetooth pulse oximeter (Medisana™ pulse oximeter PM150) and a Bluetooth digital stethoscope (3M™ Littmann® Electronic Stethoscope Model 3200) were provided and kept inside the ICU at all times. Following Kilintzis et al. (2022), the protocol included pulmonary auscultation in six locations, namely: (1) right lung-apex front, (2) right lung-base front, (3) right lung-base back, (4) left lung-apex front, (5) left lung-base front, and (6) left lung-base back, and cardiac auscultation in four locations: the standard auscultations points for the (7) aortic valve, (8) pulmonic valve, (9) mitral valve and, (10) tricuspid valve. Measurements were conducted upon the initial phase of the ICU stay and on occasions the intensive care doctors deemed necessary for clinical decision making (for example, when the patients displayed signs of clinical deterioration). Data were collected during all the discrete phases of the COVID-19 pandemic in northern Greece and included hospitalized patients affected by all the main variants of SARS-CoV2 virus, except for the Omicron strain.

In order to evaluate the effectiveness of the predictive algorithm, two different mortality rates were considered: the mortality rate within the ICU and the 90-day mortality, a commonly accepted period of time ensuring all-causes mortality is taken into account. The outcome during the ICU stay was recorded and 3 months after discharge from the ICU, survivors were contacted by phone to receive feedback on their health status at that point of time.

### 2.2. Chest X-rays pre-processing

The number of longitudinal CXRs per subject varied throughout the ICU stay and its distribution per surviving and non-surviving subjects is shown in Fig. 2. We chose 8 CXRs for each subject as it was the average value of the number of scans among all patients, meaning that if the number of CXRs was greater than 8, 8 scans would be selected randomly by intervals from all of their scans per epoch, while if the number of CXRs was fewer than 8, the variational autoencoder (VAE) would reconstruct the missing scans to reach a total of 8. All CXRs were cropped to maintain a square area around the lung field in order to remove all irrelevant annotations. This area was segmented by a U-net algorithm (Ronneberger, Fischer, & Brox, 2015), which had previously been trained on images from two publicly available datasets (Montgomery Jaeger et al., 2014 and JSRT CXRs datasets Shiraishi et al., 2000) for semantic segmentation of the lung fields. Next, all cropped images were resized to 512 × 512 and normalized between 0 and 1. Examples of 8 CXRs for two different outcomes (i.e., alive or dead) of patients are shown in Appendix Fig. 5.

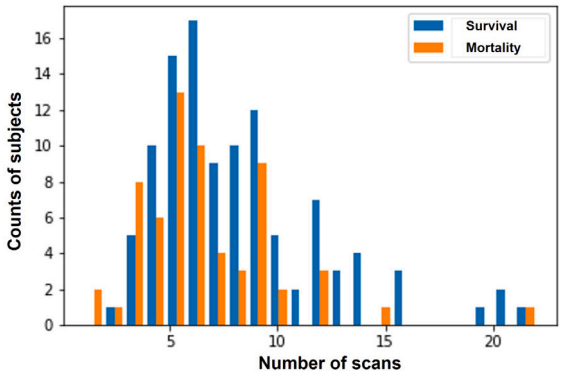


**Fig. 2.** Distribution of the number of longitudinal scans of chest X-rays per subject per class in the ICU. The number of scans ranges from 1 to 21.

### 2.3. Respiratory sounds pre-processing

A flowchart of the audio processing pipeline is shown in Fig. 3. As there were multiple audio files in each day of recordings and a single feature vector was needed for each day, the process detailed below was applied to each audio file. Two paths were followed: (i) the adventitious respiratory sounds path, and (ii) the raw audio path.

#### 2.3.1. Adventitious respiratory sounds path

In this path, the presence of adventitious respiratory sounds was detected and audio features were extracted from the detected events. The following steps were carried out:

- the signal was decomposed into intrinsic mode functions (IMFs) using empirical mode decomposition (EMD) (Huang et al., 1998);
- the first two IMFs were extracted, as most of the adventitious respiratory sounds' energy is concentrated on those IMFs;
- a Bump scalogram was obtained for each IMF waveform by computing the continuous wavelet transform (CWT) between 100 and 1600 Hz;
- the wavelet magnitude was normalized by its maximum value and a binary image was generated by applying Otsu's threshold (Otsu, 1979);
- all non-flat connected components (CCs) that had duration between 10 ms and 2 s, encompassing the typical duration of crackles, squawks, and wheezes, were selected;
- 13 Mel-frequency cepstral coefficients (MFCCs) and their deltas for each CC were computed;
- the CCs were split according to duration, i.e., those between 10–50 ms as potential crackles, those between 50–200 ms as potential squawks, and those between 100–2000 ms as potential wheezes;

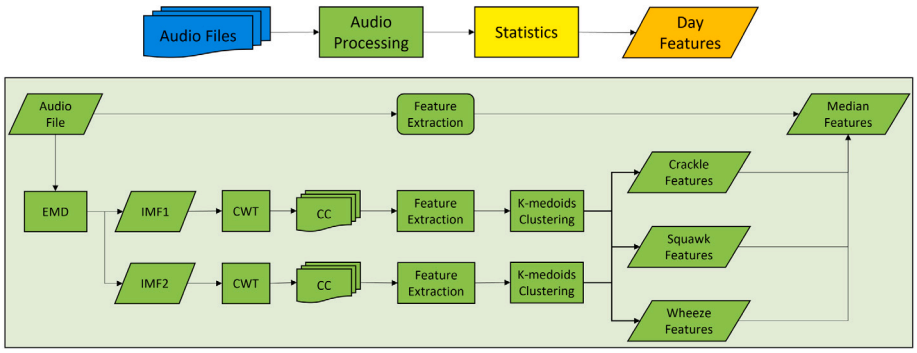


**Fig. 3.** Flowchart of the respiratory sounds pre-processing method for the audio files of a single patient on a single day. CC: connected component; CWT: continuous wavelet transform; EMD: empirical mode decomposition; IMF: intrinsic mode frequency.

- after centering the MFCC features to have median 0, the components were partitioned into clusters using the k-medoids algorithm.
- to determine the number of clusters, tests were carried out with $k = \{2, 3, \ldots, n\}$, where $n$ was the number of CCs of each duration, and the chosen value of $k$ was the one that maximized the median of the silhouette, which is a measure of how similar each point is to points in its own cluster, when compared to points in other clusters (Rousseeuw, 1987); this process was conducted for both IMFs;
- for each type of sound, the component with duration closest to the median duration of the cluster was chosen as the candidate component;
- 43 other features were extracted for the candidates, including 17 features that were previously used for squawk detection (Rocha, Pessoa, Cheimariotis et al., 2021), and 13 gammatone cepstral coefficients (GTCCs) and respective delta-GTCCs.

#### 2.3.2. Raw audio path

In this path, respiratory sounds features were computed for the total duration of each audio file. The spectrograms were computed using a Hamming sliding window with two window lengths, 32 and 64 ms, and 75% overlap. Then, 81 features previously used for wheeze and crackle classification (Rocha, Pessoa, Marques, Carvalho and Paiva, 2021) were extracted from each frame of the spectrogram: 25 spectral features, 26 MFCC features, and 30 melodic features. Most features were extracted using the MIR Toolbox 1.7.2 (Lartillot & Toiviainen, 2007). The examples of spectrograms of respiratory sounds for alive and dead subjects are shown in Appendix Fig. 6.

#### 2.3.3. Summary statistics

Table 1 provides a small description of all the respiratory sounds features from both paths. To obtain a single feature vector for each day, we computed the following summary statistics for each feature: minimum, median, maximum, mean, and standard deviation. Therefore, the total number of features at the end of the respiratory sounds pre-processing was 1440 for each window length.

### 2.4. Clinical features

A total of six clinical variables were selected, including age, daily sequential organ failure assessment (SOFA) score, Charlson comorbidity index, the oxygenation index, the ventilation equilibrium, and the dynamic compliance. The demographic distribution of the clinical features is shown in Table 2. Specifically, the conditions under which the digital auscultations took place, namely the patients' daily SOFA score, as well as the ventilation mode parameters and the arterial blood gases values of the patients were recorded, whenever available. From the ventilation mode parameters and the arterial blood gases values, the lung static compliance, the oxygenation index (i.e., the product of the Inspired Oxygen Fraction and the Mean Airway Pressure, divided by the Partial Pressure of Oxygen in the arterial blood) and a ventilation equilibrium parameter (i.e., the ratio of the minute ventilation to the partial pressure of Carbon Dioxide in the arterial blood (PaCO2) in milliliters per millimeters of mercury (ml/mmHg)) were also recorded or calculated. These clinical parameters were found to play a significant role in the description of the clinical status of the patients' respiratory system after preliminary analyses from the bulk of data obtained from the ICU files (as shown by regression analysis among numerous clinical parameters measured in the ICU). They are also in agreement with clinical parameters shown by other researchers to correlate significantly with the ICU mortality of patients with COVID-19 (Alser et al., 2021; Ferrando et al., 2020; Gallo Marin et al., 2021; Leoni et al., 2021; Xie et al., 2020).

### 2.5. Pre-trained variational autoencoder

Variational autoencoder (VAE) is a popular unsupervised learning method, which contains an encoder $\mathbf{Q}$ and a decoder $\mathbf{P}$ (Kingma & Welling, 2013), where the encoder $\mathbf{Q}_\varphi(\mathbf{Z}|\mathbf{X})$ maps the input $\mathbf{X} \in \mathbb{R}^{\mathbf{N}\times\mathbf{N}}$ to a latent representation $\mathbf{Z} \in \mathbb{R}^k$ and the decoder $\mathbf{P}_\theta(\mathbf{X}|\mathbf{Z})$ reconstructs that representation back to the input $\hat{\mathbf{X}} \in \mathbb{R}^{\mathbf{N}\times\mathbf{N}}$. $\varphi$ and $\theta$ are trainable model weights for the encoder and decoder. $\mathbf{Z}$ is the distribution over the latent space. The model is trained to minimize the reconstruction error between the input image and the reconstructed image, i.e., $\left\|\mathbf{X}-\hat{\mathbf{X}}\right\|^2$. VAE had two roles in the total methodology. Its first role was to pre-train its encoder. As the number of subjects was limited, VAE was used to extract the underlying discriminative image features, so the encoder (i.e., the layers before the bottleneck), could be fine-tuned in the next classification task rather than being trained from scratch. Its second role was to generate new synthetic images by learning the distribution of the recorded images and decoding from that distribution. If the number of CXRs was fewer than 8, VAE would reconstruct the last scan of that patient to generate enough scans. For example, if a subject had 5 scans, the VAE model would generate 3 more scans based on that subject's 5th scan. Since the distribution of $\mathbf{Z}$ was different each time, the new reconstructed images were slightly different from each other. In addition, in order to make $\mathbf{Q}_\varphi(\mathbf{Z}|\mathbf{X})$ to be as close as possible to $\mathbf{P}_\theta(\mathbf{X}|\mathbf{Z})$, the Kullback–Leibler (KL) divergence was used to measure how similar the two distributions were. Therefore, the total loss function $\mathcal{L}$ of VAE model was computed as:

$$\mathcal{L} = \gamma \cdot \mathbf{KL}\left[\mathbf{Q}_{\mathbf{Z}|\mathbf{X}} \parallel N(\mathbf{0},\mathbf{I})\right] + \left\|\mathbf{X}-\hat{\mathbf{X}}\right\|^2, \tag{1}$$

where KL divergence calculates the similarity between the posterior distribution $\mathbf{Q}_{\mathbf{Z}|\mathbf{X}}$ and the standard Gaussian distribution $N(\mathbf{0},\mathbf{I})$. $\gamma$ is a hyperparameter that controls the importance of these two terms. Overall, the encoder network $\mathbf{Q}_\varphi(\mathbf{Z}|\mathbf{X})$ maps input $\mathbf{X}$ to latent $\mathbf{Z}$, where $\mathbf{Z}$ is made up of two parameters, a mean vector $\mu \in \mathbb{R}^k$ and a standard

**Table 1**
Small description of the respiratory sounds features.

| Path | Feature | Description of each extracted feature |
|---|---|---|
| Adventitious respiratory sounds path | Duration | Duration of event |
| | Fundamental frequency | Minimum frequency |
| | Frequency range | Frequency range |
| | Zero-crossing rate | Number of zero-crossings per second |
| | IMF1 peaks | Number of peaks above 1/4 of the maximum amplitude of IMF1 of each event |
| | Graphical extent | Ratio of pixels in the CC to pixels in the total bounding box |
| | Graphical perimeter area | Ratio of pixels around the boundary of the CC to pixels in the CC |
| | Spectral centroid | Center of mass of the spectral distribution |
| | Spectral crest | Ratio between the maximum spectral value and the arithmetic mean of the energy spectrum value (Peeters, 2004) |
| | Spectral entropy | Estimation of the complexity of the spectrum |
| | Spectral flatness | Estimation of the noisiness of a spectrum |
| | Spectral kurtosis | Measure of the flatness of a distribution around its mean value |
| | Spectral rolloff | Frequency such that 95% of the total energy is contained below it |
| | Spectral skewness | Measure of the asymmetry of a distribution around its mean value |
| | Spectral slope | Linear regression of the magnitude spectrum |
| | Spectral spread | Variance of the spectral distribution (Lerch, 2012) |
| | Harmonic ratio | Maximum of the normalized autocorrelation |
| | MFCC | 13 Mel-frequency cepstral coefficients |
| | Delta-MFCC | 1st-order temporal differentiation of the MFCCs |
| | GTCC | 13 Gammatone cepstral coefficients |
| | Delta-GTCC | 1st-order temporal differentiation of the GTCCs |
| Raw audio path | Spectral centroid | Center of mass of the spectral distribution |
| | Spectral spread | Variance of the spectral distribution |
| | Spectral skewness | Skewness of the spectral distribution |
| | Spectral kurtosis | Excess kurtosis of the spectral distribution |
| | Zero-crossing rate | Waveform sign-change rate |
| | Spectral entropy | Estimation of the complexity of the spectrum |
| | Spectral flatness | Estimation of the noisiness of a spectrum |
| | Spectral roughness | Estimation of the sensory dissonance |
| | Spectral irregularity | Estimation of the spectral peaks' variability |
| | Spectral flux | Euclidean distance between the spectrum of successive frames |
| | Spectral flux Inc | Spectral flux with focus on increasing energy solely |
| | Spectral flux halfwave | Halfwave rectified spectral flux |
| | Spectral flux median | Median filtered spectral flux |
| | Spectral brightness | Amount of energy above 100, 200, 400, and 800 Hz |
| | Brightness 400 ratio | Ratio between spectral brightness at 400 and 100 Hz |
| | Brightness 800 ratio | Ratio between spectral brightness at 800 and 100 Hz |
| | Spectral rolloff | Frequency such that 95, 75, 25, and 5% of the total energy is contained below it |
| | Rolloff outlier ratio | Ratio between spectral rolloff at 5 and 95% |
| | Rolloff interquartile ratio | Ratio between spectral rolloff at 25 and 75% |
| | MFCC | 13 Mel-frequency cepstral coefficients |
| | Delta-MFCC | 1st-order temporal differentiation of the MFCCs |
| | Pitch | Fundamental frequency estimation |
| | Pitch smoothing | Moving average of the pitch curve with lengths of 100, 250, 500, and 1000 ms |
| | Inharmonicity | Partials non-multiple of fundamental frequency |
| | Inharmonicity smoothing | Moving average of the inharmonicity curve with lengths of 100, 250, 500, and 1000 ms |
| | Voicing | Presence of fundamental frequency |
| | Voicing smoothing | Moving average of the voicing curve with lengths of 100, 250, 500, and 1000 ms |

**Table 2**
Demographic distribution of the clinical variables.

| Features | Mean (±std) |
|---|---|
| Number of subjects | 171 |
| Age | 65.40 ± 10.16 |
| Sequential organ failure assessment | 6.01 ± 2.65 |
| Charlson comorbidity index | 3.60 ± 2.00 |
| Oxygenation index | 0.90 ± 0.27 |
| Ventilation equilibrium | 0.47 ± 0.50 |
| Dynamic compliance | 44.64 ± 45.03 |

deviation vector $\sigma \in \mathbb{R}^k$, so the decoder can sample from these two distributions to reconstruct $\mathbf{Z}$ back to $\hat{\mathbf{X}}$. $\mathbf{Z} = \mu + e^{\sigma} \cdot \epsilon$, where $\epsilon$ is a random normal tensor. In this work, the parameter values were the following: $\mathbf{N} = 512$, $k = 128$ and $\gamma = 0.01$, which were selected from ablation studies.

Specifically, as shown in Fig. 4A, the architecture of VAE model is non-symmetric (He et al., 2021). ResNet-50 (He, Zhang, Ren, & Sun, 2016) was selected as the encoder of VAE to extract highly discriminative image features. Previous studies focusing on Covid-19 detection using CXRs have demonstrated superior performance with ResNet-50 compared to other architectures (Narin, Kaya, & Pamuk, 2021). Its deep architecture, complemented by residual connections, addresses the challenges of training deep neural networks and mitigates the problem of vanishing gradients. In addition, the global average pooling layer (GAP) was added to reduce the dimensionality of the feature maps and to better represent the latent vectors. As the goal was to use the pre-trained encoder to produce image representations for the subsequent classification task, the decoder was designed to be lightweight and shallow, which significantly reduced the training computation and time (He et al., 2021). Next, a fully-connected (FC) layer was used to change the dimension of latent vectors to be the same as the GAP layer and reshape it to $16 \times 16 \times 32$, followed by five convolutional transpose layers (Conv2DT) that upsampled the feature maps to the shape of original images. Each Conv2DT layer (except the last layer) had 64 kernels with size of $3 \times 3$ and stripes of 2, followed by the exponential linear unit (ELU) activation function. ELU (Clevert, Unterthiner, & Hochreiter, 2015) was chosen because it is continuous and differentiable at all points and avoids the "dying ReLU" problem. The last Conv2DT layer had only 1 kernel with the same size and stripes as before. Other implementation details included that the batch size was 4, the Adam was chosen as the optimizer with an initial learning

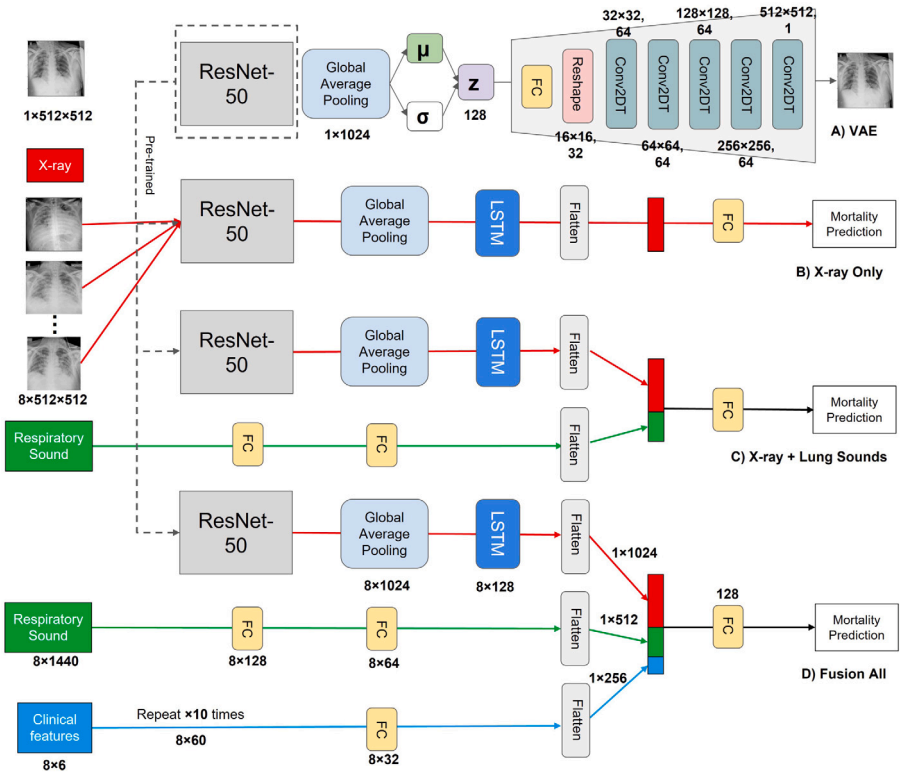


**Fig. 4.** The pipeline of the model to make mortality predictions on patients with COVID-19 in the ICU. A: the variational autoencoder (VAE) reconstructs each CXR to pre-train the encoder (i.e., ResNet-50). B: the model architecture that uses only longitudinal CXRs as the input. C: the model architecture that fuses CXRs and respiratory sounds features as the input. D: the proposed model architecture that fuses CXRs, respiratory sounds features and selected clinical features as the input.

rate of 0.0001 and the early stopping with a patience of 8 was applied to prevent overfitting. All real and synthetic CXRs were involved and were split into training (80%) and validation sets (20%) based on the number of subjects. As a result, after the VAE was fully trained, the pre-trained encoder was applied to the fusion model in the later classification task and the decoder was used to generate reconstructed images.

### 2.6. The sequential fusion model

The proposed fusion model has three branches, i.e., the CXRs branch, the respiratory sounds branch and the clinical features branch. The features from each branch are fused together to predict the mortality of patients with COVID-19 in the ICU or 90 days after discharge from the ICU, as shown in Fig. 4.

The CXRs branch took 8 longitudinal images ($8 \times 512 \times 512$) as the input to ResNet-50, which had been pre-trained in the previous stage. Similarly, the GAP layer reduced the dimensionality of the feature map to a 2048-dimensional feature vector for each CXR, i.e., in $8 \times 2048$, which was then fed into the Long Short Term Memory (LSTM) network. LSTM is a unique type of recurrent neural networks capable of handling long-term dependencies. It has memorial cells to maintain its cell status over time and different types of gates to optionally add or remove the information from cells. Therefore, these two components controlled how much information of the CXR at this timestamp could be passed to the CXR at the next timestamp. LSTM was used to extract the sequential features of CXRs over time for each subject. The size of all the gates was 128, so the output feature size of LSTM was $8 \times 128$. After the flattening layer, the feature size from the CXRs branch was 1024.

The respiratory sounds branch took the pre-processed respiratory sounds features ($8 \times 1440$) as input, which were fed into two fully connected layers with 128 and 64 neurons. After the flattening layer, the feature size from the respiratory sounds branch was 512.

The clinical features branch took the selected clinical features as input. In order to extract useful features from such small feature size, the clinical features were not directly fed into the model but were repeated ten times for each input ($8 \times 60$). One fully connected layer with 32 neurons was added to the model and after the flattening layer, the feature size from the clinical features was 256.

Furthermore, the features from these three branches were concatenated together in size of 1792 and were fed into two fully connected layers with 128 neurons and 1 output neuron, followed by a *sigmoid* activation function to output a mortality probability ranging from 0 to 1.

The loss function of the fusion model, $\mathcal{L}_f$, is the weighted binary cross-entropy, which was defined as:

$$\mathcal{L}_f = -\frac{1}{N}\sum_{i=1}^{N}\left\{\lambda \cdot y_i \cdot log(\hat{y}_i) + (1-\lambda)\cdot(1-y_i)\cdot log(1-\hat{y}_i)\right\}, \quad (2)$$

where $y_i$ is the ground truth label for the mortality prediction, where 0 stands for the subject is survival while 1 stands for the subject is not survival. $\hat{y}_i$ is the predicted probability from the model and $N$ is the batch size. $\lambda$ is a hyper-parameter that weights the loss function to overcome the imbalanced dataset, which is represented by calculating the frequency of two classes and then inverting them so that the underrepresented class has a much higher error than the majority class. Other implementation details were that models were trained with Adam optimizer with an initial learning of 0.0005, the batch size was 4 per step and the early stopping was with patience of 8.

## 3. Results

### 3.1. Experimental design

A collection of 1271 CXRs acquired from 171 subjects (114 men, 57 women, 65.4 ± 10.16 years-old) were included in this study. The

**Table 3**
Comparisons of model performances using different inputs for ICU mortality prediction and 90 days mortality prediction.

| Results for 90 days mortality (Mean ± Std) | | | | | |
|---|---|---|---|---|---|
| Model | Accuracy | Precision | Recall | F1 | AUC |
| CXRs only | 0.703 ± 0.008 | 0.723 ± 0.017 | 0.728 ± 0.034 | 0.721 ± 0.020 | 0.709 ± 0.019 |
| CXRs + Respiratory sounds | 0.738 ± 0.010 | 0.752 ± 0.018 | 0.743 ± 0.028 | 0.743 ± 0.017 | 0.742 ± 0.011 |
| Fusion all | **0.743 ± 0.012*** | **0.755 ± 0.025*** | **0.750 ± 0.029*** | **0.750 ± 0.022*** | **0.752 ± 0.012*** |
| Results for ICU mortality (Mean ± Std) | | | | | |
| Model | Accuracy | Precision | Recall | F1 | AUC |
| CXRs only | 0.715 ± 0.006 | 0.724 ± 0.020 | 0.732 ± 0.022 | 0.724 ± 0.018 | 0.718 ± 0.014 |
| CXRs + Respiratory sounds | 0.754 ± 0.014 | **0.760 ± 0.023*** | 0.751 ± 0.019 | 0.752 ± 0.022 | 0.751 ± 0.009 |
| Fusion all | **0.761 ± 0.011*** | 0.758 ± 0.027 | **0.782 ± 0.024*** | **0.766 ± 0.019*** | **0.759 ± 0.008*** |

*Denotes the comparison is statistically significant ($p < 0.05$) between this result with any of the other results.
The best results are highlighted in bold.

distribution of the number of CXRs per subject is shown in Fig. 2. Two binary outcomes of the mortality of patients with COVID-19 were predicted, i.e., the ICU mortality and the 90 days mortality. For the ICU mortality prediction, there were 63 positive subjects (i.e., dead) and 108 negative subjects (i.e., alive), and for the 90 days mortality, there were 80 positive subjects and 91 negative subjects. We used four-fold cross-validation for each experiment, and the model ran five independent times for each fold to avoid overfitting. Specifically, the subjects were split into 75% as training and validation sets, and 25% as the separate testing set. In order to justify the effectiveness of the fusion model, we ran additional experiments on the CXRs dataset only and CXRs with respiratory sounds features. Overall, six models were evaluated: **(1)** 90 days mortality on CXRs only; **(2)** 90 days mortality on CXRs and respiratory sounds features; **(3)** 90 days mortality on CXRs, respiratory sounds features and clinical features; **(4)** ICU mortality on CXRs only; **(5)** ICU mortality on CXRs and respiratory sounds features; **(6)** ICU mortality on CXRs, respiratory sounds features and clinical features. In addition, we ran several more experiments to compare the model performances among the choices of longitudinal days and the window lengths used to extract respiratory sounds features (i.e., 32 ms or 64 ms). Regarding evaluation measures, accuracy, precision, recall, and F1 score were used to evaluate model performances and the threshold was set as 0.5. The AUC score was calculated for this binary classification task at various threshold settings. The overall mean and standard deviation values were calculated for each metric. In addition, the paired t-test was performed to compare different model results. The comparison was considered as statistically significant if $p < 0.05$. All experiments and statistical analyses were performed using two GPUs (Nvidia Quadro RTX 8000) with Tensorflow 2.7 and Scikit-learn packages in Python 3.7.

### 3.2. Evaluation of VAE

In this paper, the VAE was employed to pre-train the encoder for subsequent use as a feature extractor in the classification task, as well as to generate synthetic images. To assess the performance of the VAE, we calculated several metrics on the test set, including the mean squared error (MSE), the structural similarity index (SSIM), and the peak signal-to-noise ratio (PSNR). These metrics were utilized to evaluate the similarity between the generated CXRs and the ground truth CXRs. Specifically, the overall MSE score is $0.090 \pm 0.007$, the SSIM score is $0.697 \pm 0.052$ and the PSNR score is $21.683 \pm 3.317$. The generated missing CXRs are shown in Appendix Fig. 7.

### 3.3. Model comparisons

The main results are shown in Table 3 for ICU mortality prediction and 90-day mortality prediction. For each prediction, three models were compared using different input modalities, including CXRs, CXRs and respiratory sounds features, and CXRs, respiratory sounds features, and clinical features. Overall, the models' performance on the ICU mortality prediction task outperformed the performance on the 90-day mortality prediction task for all metrics. For ICU mortality prediction, the model fusing all features achieved an AUC of 0.759, an accuracy of 0.761, a precision of 0.758, a recall of 0.782 and a F1 score of 0.766. Although the precision score (0.758) was slightly worse than that of the model with CXRs and respiratory sounds (0.760) as the input, fusing all three types of features significantly improved the model's performance among other metrics ($p < 0.05$). Moreover, the model fusing respiratory sounds features and CXRs achieved a better performance than the one containing only CXRs, improving the accuracy from 0.715 to 0.754 and the AUC from 0.718 to 0.751. Furthermore, in order to validate our findings, we employed t-distributed stochastic neighbor embedding (t-SNE) plots to reduce the dimensionality of the feature vectors at the Fully Connected layers to two, as illustrated in Fig. 4. By visualizing their distributions, as shown in Appendix Fig. 8, it is evident that the feature distributions achieved through the fusion of all three modalities exhibit superior separation compared to the fusion of only two modalities (CXR and respiratory sounds) or utilizing CXRs alone. Similarly, for the 90-day mortality prediction, the model fusing all features outperformed the other two models ($p < 0.05$) and the model with respiratory sounds and CXRs achieved better performance than the CXRs only model. Also, compared to simpler single-point severity metrics, like SOFA and APACHE scores, the proposed fusion model yielded better results, especially at the long-term prognosis field, since it was able to provide reliable predictions for the 90-day survival.

Additionally, further analysis was performed on specific cases that were misclassified by the proposed algorithm (either as false survival or false death in the ICU) in order to detect possible common clinical patterns in these cases. The post-hoc analysis revealed that, in 4 cases where patients were misclassified as survivors in the ICU, only one auscultation session was performed (in 3 cases) and the radiologic findings were characterized as limited compared to more severe cases, which was likely to lead to the misclassifications. In two occasions, the patients perished after acute complications during their stay in the ICU (mainly barotrauma cases). On the other hand, the falsely classified death cases were five, and three of those patients eventually died within 90 days after the discharge from the ICU. Also, in four cases, radiologic findings were indicative of very severe ARDS bilaterally, whereas the quality of the obtained auscultation files was poor, hampering the validity of the feedback for our algorithmic solution.

### 3.4. Model fine-tuning

In order to achieve the optimal performance of the model, several ablation studies were conducted, including the number of longitudinal CXRs each subject required and the window lengths for pre-processing respiratory sounds features. All details are included in Tables 4 and 5.

The number of longitudinal CXRs and other features that the model needed to predict the mortality were examined. The effects of using recordings of 3 days, 8 days, and 13 days were compared. In terms of mortality predictions, especially for the urgent cases, such as the

**Table 4**
Ablation studies on the choice of the longitudinal days.

| Predictions for the 90 days mortality using different longitudinal days (Mean ± Std) | | | | |
|---|---|---|---|---|
| Longitudinal days | Accuracy | Precision | Recall | AUC |
| 3 days | 0.551 ± 0.031 | 0.559 ± 0.033 | 0.547 ± 0.040 | 0.593 ± 0.025 |
| 8 days | **0.743 ± 0.012*** | **0.755 ± 0.025*** | **0.750 ± 0.029*** | **0.752 ± 0.012*** |
| 13 days | 0.701 ± 0.018 | 0.723 ± 0.020 | 0.712 ± 0.033 | 0.721 ± 0.017 |
| Predictions for the ICU mortality using different longitudinal days (Mean ± Std) | | | | |
| Longitudinal days | Accuracy | Precision | Recall | AUC |
| 3 days | 0.593±0.025 | 0.602 ± 0.016 | 0.591 ± 0.019 | 0.609 ± 0.011 |
| 8 days | **0.761 ± 0.011*** | **0.758 ± 0.027*** | **0.782 ± 0.024*** | **0.759 ± 0.008*** |
| 13 days | 0.709 ± 0.020 | 0.731 ± 0.015 | 0.725 ± 0.020 | 0.730 ± 0.012 |

*Denotes the comparison is statistically significant ($p < 0.05$) between this result with any of the other results. The best results are highlighted in bold.

**Table 5**
Ablation studies on different window lengths of lung sounds (32 ms and 64 ms).

| Predictions for the ICU mortality using different window lengths of respiratory sounds (Mean ± Std) | | | | |
|---|---|---|---|---|
| Window length | Accuracy | Precision | Recall | AUC |
| CXRs + Respiratory sounds (32 ms) | 0.748 ± 0.012 | 0.753 ± 0.021 | 0.743 ± 0.020 | 0.744 ± 0.011 |
| CXRs + Respiratory sounds (64 ms) | 0.754 ± 0.014 | **0.760 ± 0.023*** | 0.751 ± 0.019 | 0.751 ± 0.009 |
| Fusion all (32 ms) | 0.758 ± 0.008 | 0.751 ± 0.020 | 0.774 ± 0.019 | 0.753 ± 0.007 |
| Fusion all (64 ms) | **0.761 ± 0.011*** | 0.758 ± 0.027 | **0.782 ± 0.024*** | **0.759 ± 0.008*** |

*Denotes the comparison is statistically significant ($p < 0.05$) between this result with any of the other results.
The best results are highlighted in bold.

subjects in the ICU, it is important to use the least time to make the correct prognosis and treatment for them. Therefore, it was examined if the model could achieve a comparable result by only using the first three-days features from CXRs, respiratory sounds, and clinical features. All pre-processing steps remained the same, but only the first three scans were chosen for each subject. The results were underwhelming, i.e., for ICU mortality prediction, the model only achieved an accuracy of 0.593, a precision of 0.602, a recall of 0.591 and an AUC of 0.609. The performance of the 90-day mortality prediction was worse than the ICU mortality prediction, which was consistent with our previous findings, with the model achieving an accuracy of 0.551, a precision of 0.559. a recall of 0.547 and an AUC of 0.593. Next, it was examined if more longitudinal CXRs could achieve a better mortality prediction, so we conducted another experiment on 13 days. However, the results showed that more CXRs and features did not guarantee a better model performance. For the ICU mortality prediction, the model achieved an accuracy of 0.709, a precision of 0.731, a recall of 0.725 and an AUC of 0.730, and for the 90-day prediction, the model achieved an accuracy of 0.701, a precision of 0.723, a recall of 0.712 and an AUC of 0.721. They were both worse than the best models on 8 days.

Furthermore, another ablation study was performed to understand the impact on ICU mortality prediction of the window lengths (i.e., 32 ms or 64 ms) that were used to process respiratory sounds. The results showed that models trained with respiratory sounds features extracted using a 64 ms window length achieved a slightly better performance than those that used a window length of 32 ms. For models trained on the fusion of CXRs and respiratory sounds features, the 64 ms window length improved the AUC from 0.744 to 0.751 and for models trained on all three types of features, the 64 ms window length improved the AUC from 0.753 to 0.759.

## 4. Discussion

In this study, we proposed a deep fusion model that utilized longitudinal CXRs, respiratory sounds features and clinical features to predict the mortality of patients with COVID-19 in the ICU. By comparing results with the CXRs only model, it was observed that the addition of respiratory sound features and clinical features significantly improved the mortality prediction, achieving an accuracy of 0.761, a precision of 0.758, a recall of 0.782, a F1 score of 0.766 and an AUC of 0.759. Moreover, the significant improvement in the performance of the fusion model compared to the other models suggests that the fusion model has the potential to better predict the mortality of critically ill ICU patients with COVID-19, which may facilitate the decision-making process and improve clinical triage systems.

Another interesting finding of this paper is that the fusion model has the potential for longer-term mortality prediction. We used the same dataset for two predictions, i.e., ICU mortality prediction and 90-day mortality prediction. As expected, we see that the model performances were better in the former task than in the latter in Table 3, because all the input data were collected during their ICU period. There were several patients that survived the ICU but died in the following 90 days, so their ground truth was different in the two tasks. We found that the fusion model for ICU mortality predictions incorrectly classified most of these cases as non-surviving. Although these patients were still alive during their stay in the ICU, their status changed after 90 days as predicted by our model. This indicates that our fusion model was able to use the data collected during the ICU stay for a long-term mortality prediction. Moreover, as shown in Table 4, we tried to see if the model could make a prediction using the data collected in the first three days. However, the results were disappointing. Possible reasons are the complexity of the severely ill COVID-19 cases, with multiple clinical parameters and organ failures occurring and interacting with each other over the course of several days during the ICU stay. Previous studies have consistently supported this finding, demonstrating significant associations between progressive imaging patterns indicative of lung abnormality and mortality (Putman et al., 2019). By capturing the evolving nature of lung abnormalities over time, the variation between CXR slices can offer additional insights into the patient's condition. This could also explain the apparently lower accuracy of the model in cases where only single measurements were obtained. Likewise, the models in Table 3 show some misclassifications, possibly due to the patients not having a relatively long sequential dataset. These observations show that a limitation of this study is that the model requires a longer sequence data to guarantee better model performance.

This study has a number of strengths that need to be acknowledged. First, the encoder was pre-trained with a VAE model at the initial stage, which had a dual positive effect. On the one hand, due to the limited cases in this study, training a deep neural network on such a limited dataset can lead to the problem of overfitting. Therefore, as an unsupervised learning algorithm, VAE was able to extract underlying image

**Table 6**
Comparisons between the proposed model and state of the art algorithms.

| Paper | Task | Data | Method | Performance |
|---|---|---|---|---|
| Bae et al. (2021) | Mortality risk prediction | CXRs + clinical variables | ResNet50 + RF/LDA | AUC = 0.83 |
| Shamout et al. (2021) | Deterioration risk prediction | CXRs + clinical variables | Deep CNNs | AUC = 0.786 |
| Kwon et al. (2021) | Mortality risk prediction | CXRs + clinical variables | DenseNet-121 | AUC = 0.82, Acc = 0.42, Pre = 0.27, Rec = 0.78, F1 = 0.41 |
| Aljouie et al. (2021) | Mortality risk prediction | CXRs + demographic and clinical variables | SVM, RF, LR, XGB | AUC = 0.83, Rec = 1, Spe = 0.61 |
| Gourdeau et al. (2022) | Mechanical ventilation in ICU | Pre-intubation CXRs + risk factors | DenseNet-121 | AUC = 0.743, Acc = 0.755, Rec = 0.487, Spe = 0.828 |
| Cheng et al. (2022) | Mortality risk prediction in ICU | CXRs + clinical variables | Transformer-based CNNs | AUC = 0.727, Acc = 0.732, Rec = 0.714, Spe = 0.746, F1 = 0.707 |
| Our work | Mortality risk prediction in ICU | CXRs + respiratory sounds + clinical variables | VAE + fusion model | AUC = 0.759, Acc = 0.761, Pre = 0.758, Rec = 0.782, F1 = 0.770 |

CXR: chest X-ray; RF: random forest; LDA: linear discriminant analysis; CNN: convolutional neural network
SVM: support vector machines; LR: logistic regression; XGB: extreme gradient boosting;
VAE: variational autoencoder; Acc: accuracy; Pre: precision; Rec: recall; Spe: specificity.

features in the latent space by trying to reconstruct the original images without additional labels. Then, the pre-trained weights of the encoder could be further initialized in the next fusion model, eliminating the need to train the deep model from scratch. Some other studies have demonstrated the effectiveness of VAE in limited datasets similar to ours. For example, Akrami et al. developed a robust VAE to detect brain lesions on a small MRI dataset and they found that the accuracy of lesion detection could be improved by first pre-training parts of the network within the VAE (Akrami, Joshi, Li, Aydore, & Leahy, 2020). On the other hand, VAE was used to generate missing CXRs. As shown in Fig. 2, the number of CXRs for subjects is quite imbalanced. To ensure that each subject had 8 CXRs as input, VAE compensated for the lack of scans by generating new images decoded from the feature distributions in the latent space. Therefore, VAE was considered a prerequisite for successfully training the fusion model.

Additionally, this is the first study that incorporated longitudinal CXRs, respiratory sounds features, and clinical features into one deep fusion model to predict the mortality of patients with COVID-19 in the ICU. The results in Table 3 demonstrate that respiratory sounds and clinical variables were effective in improving mortality prediction in severe patients with COVID-19. Longitudinal CXRs were chosen instead of a single scan because a previous study has shown that longer time-series information is able to track progressive lung severity over time, thereby improving model performances (Cheng et al., 2022). In addition, LSTM networks are able to learn long-term dependencies to extract longitudinal features of CXRs over time. Using CXRs alone to predict mortality is challenging, but several previous studies have used CXRs to predict severity in patients with COVID-19. For example, Aboutalebi et al. proposed CXR-S, a deep network for predicting the airspace severity in patients with COVID-19 on a single CXR image (Aboutalebi et al., 2021). Cohen et al. first pre-trained DenseNet on a non-COVID19 dataset, and then fine-tuned the network on their COVID-19 CXRs to predict lung opacity scores, achieving a correlation of 0.78 (Cohen et al., 2020). Recent studies have demonstrated that respiratory sounds are a reliable marker of COVID-19, as respiratory sounds (e.g., crackles) vary continuously from mild to severe patients with COVID-19 (Noda et al., 2020; Wang et al., 2020). However, only a few works collected respiratory sounds data acquired from patients with COVID-19. Pancaldi et al. detected patients with COVID-19 from velcro-like respiratory sounds by processing and extracting the characteristics of respiratory sounds using the software VECTOR (Pancaldi et al., 2022). In addition, Sait et al. developed a multi-model system that used both respiratory sounds and CXRs to diagnose patients with COVID-19, reaching an accuracy of 0.8 for respiratory sounds analysis and 0.99 for the CXRs dataset (Sait et al., 2021). Importantly, we collected longitudinal respiratory sounds in this study, and the results from Table 3 show a significant improvement in mortality predictions after fusing respiratory sounds features. Furthermore, clinical variables were proved to be effective in mortality prediction of patients with COVID-19 (Aljouie et al., 2021; Kwon et al., 2021). Based on our preliminary study, as well as on previous works on clinical features correlating with ICU mortality, we selected six clinical variables and the results show the model fusing all those features outperforms other models. The importance of the selected clinical variables is the comprehensive description of the clinical status of the lungs (both anatomically and functionally), as they include vital information on the lungs' compliance, the gas exchange status, the aeration of the lung parenchyma, and the intensity of the mechanical ventilation. Moreover, they include data concerning the presence of comorbidities and the vital organs' failure that may co-exist. The selected bio-parameters are in conjunction with reported clinical parameters that other researchers have found to be associated with ICU mortality in patients with COVID-19 (Alser et al., 2021; Ferrando et al., 2020; Gallo Marin et al., 2021; Leoni et al., 2021), strengthening the validity of their selection for inclusion in our model.

We further compared our method with other papers using different datasets in COVID-19 mortality prediction, as shown in Table 6. Different models were evaluated with different metrics, so we chose the AUC to compare the performance of all models. Although all previous methods used CXRs and clinical variables as the input, since we are the first to incorporate respiratory sounds features into the model, it is challenging to train their models on our dataset for a direct comparison here. However, from the results, we can still see that the AUCs for this mortality prediction range from 0.72 to 0.83. In addition, it is interesting to find that the overall performances on subjects in the ICU (Cheng et al., 2022; Gourdeau et al., 2022) are worse than those in the general population because patients in the ICU are by default critically ill, so it is more difficult for models to predict their outcome. We found only two papers that collected data from patients with COVID-19 in the ICU similar to our approach. Cheng et al. who used a transformer-based CNN to extract additional features from longitudinal CXRs and normal fully-connected layers on clinical variables, achieved an accuracy of 0.732 and an AUC of 0.727 (Cheng et al., 2022). Gourdeau et al. fine-tuned a pre-trained Densenet-121 network on their dataset combining single-day CXRs and selected risk factors, reaching an accuracy of 0.75 and an AUC of 0.74. However, these studies only had single-day clinical variables, whereas we had clinical variables at different time points to track informative changes during the patients' ICU stay. Moreover, most of their datasets were collected in 2020 and Cheng et al. (2022)

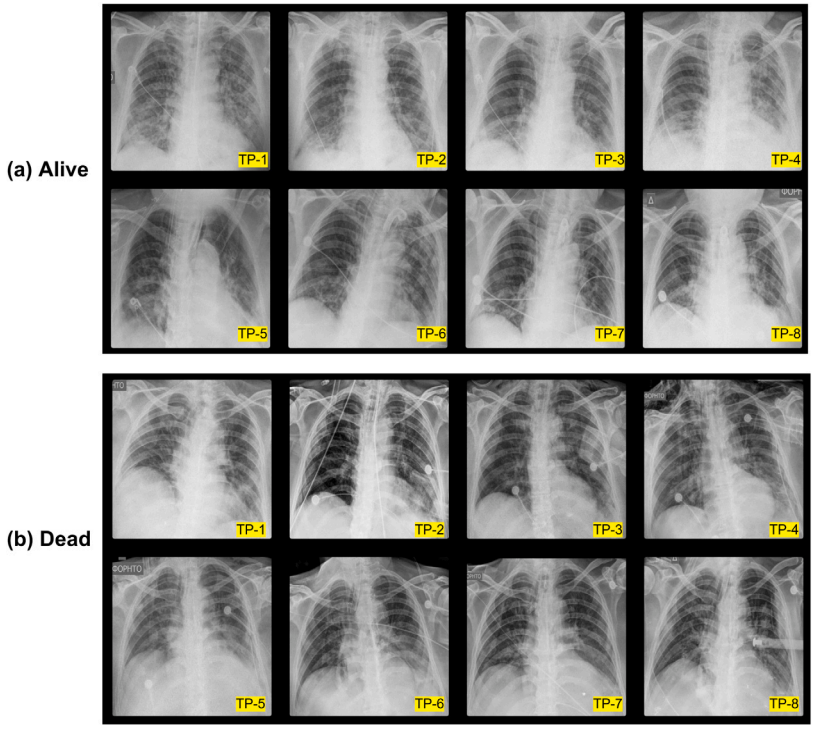


**Fig. 5.** Examples of 8 sequential CXRs from two different outcomes of patients. TP: time point. (a) The CXRs from a patient alive in the ICU; (b) The CXRs from a patient dead in the ICU.

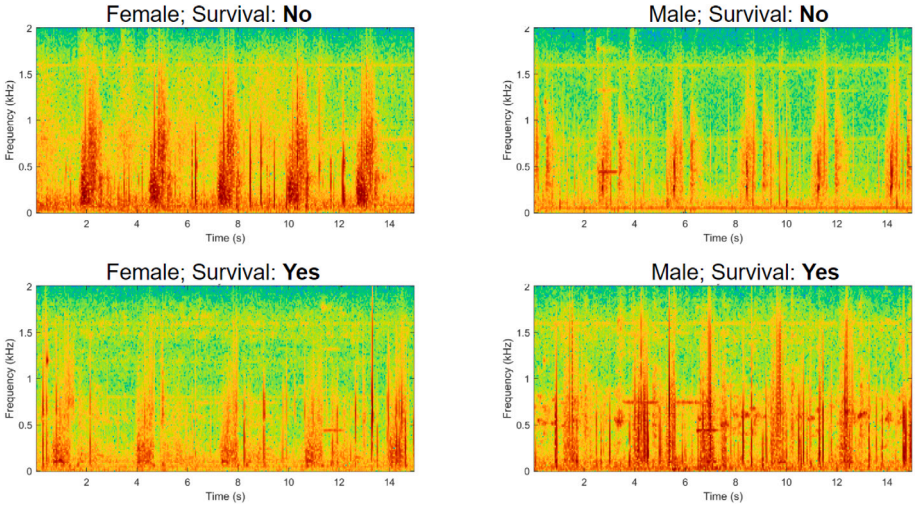


**Fig. 6.** Spectrograms of respiratory sounds for alive and dead samples.

mentioned that their cohort was entirely unvaccinated, so there is the possibility that these models would not be equally effective on fully vaccinated patients, or have altered accuracy in the subsequent SARS-CoV2 variants. The fact that our sample included patients from all the major viral variants is one of the strengths of this study, especially when one considers the inclusion of respiratory sounds from critically ill COVID-19 patients for the first time in literature.

The added value of this model to the clinical practice in critically ill patients with COVID-19 lies in the early identification of patients at risk of severe complications and subsequent adverse outcome, based on reliable and meaningful clinical data, and data derived from auscultation and radiology images from the thorax. These measurements provide continuous feedback on the underlying pathophysiology of the severely ill patients with ARDS and cover various areas of interest including the functional imaging and clinical information on the status of the affected lungs. Early identification of patients at risk in the ICU environment could facilitate automatic generation of alerts for this pool of patients enabling targeted interventions in an effort to reverse the predicted outcome. To the best of our knowledge, this is the first reported creation of a reliable prediction model that takes into consideration clinical, sound, and imaging data from severely ill patients with COVID-19.

This study has several limitations. First, the scale of the dataset in the experiment is still relatively small although we are not aware of the existence of analogous databases to date. Despite the practical difficulties to collect sequential datasets, they are expected to improve the performance of the proposed model. However, a sufficient sequential dataset is required for better classification results. As shown in Table 4, using only three days of data did not yield the expected results. In

**Fig. 7.** Examples of the generated missing CXRs for three subjects.

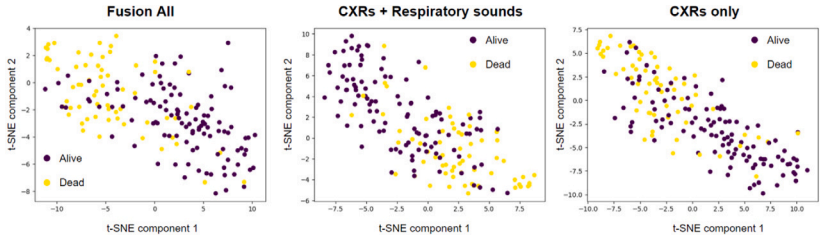


**Fig. 8.** Feature distribution of ICU mortality predictions using different modalities: (i) Fusion All, (ii) CXRs + Respiratory Sounds, and (iii) CXRs Only. The feature vectors are reduced from 128 to 2 dimensions using t-distributed stochastic neighbor embedding (t-SNE).

addition, new variants of the virus continue to emerge through mutations, making prognosis more challenging. It is our intention to collect more data from different centers in the future to further improve the generalization ability of the model. Second, the clinical features were manually selected from our preliminary experiments, so it would be interesting to figure out how the model automatically extracts features if all collected clinical information is actually available. Finally, the fusion model can be optimized. Recently, transformer-based models have been proposed to better address the time series prognosis task, which extract useful correlations in long sequences by using self-attention modules. Therefore, a future step is to incorporate transformers into our fusion model.

In conclusion, we propose a deep fusion model that predicts the mortality of patients with COVID-19 in the ICU by using sequential data from chest X-rays, respiratory sounds features, and clinical variables. The results show that the addition of respiratory sounds and clinical variables significantly improves the mortality prediction of critically ill patients in the ICU. In addition, comparisons between ICU survival predictions and 90-day survival predictions suggest that the fusion model has the potential to make successful longer mortality predictions. It provides an empowered tool to speed up clinical decision processes and save more patients' lives.

## CRediT authorship contribution statement

**Yunan Wu:** Writing – original draft, Conceptualization, Methodology, Software, Visualization, Validation. **Bruno Machado Rocha:** Writing – original draft, Software, Data curation. **Evangelos Kaimakamis:** Resources, Conceptualization, Investigation, Writing – review & editing. **Grigorios-Aris Cheimariotis:** Data curation, Writing – review & editing. **Georgios Petmezas:** Data curation, Writing – review & editing. **Evangelos Chatzis:** Investigation, Writing – review & editing. **Vassilis Kilintzis:** Resources, Writing – review & editing. **Leandros Stefanopoulos:** Investigation, Writing – review & editing. **Diogo Pessoa:** Data curation, Resources, Writing – review & editing. **Alda Marques:** Writing – review & editing, Conceptualization. **Paulo Carvalho:** Writing – review & editing, Conceptualization. **Rui Pedro Paiva:** Writing – review & editing, Investigation. **Serafeim Kotoulas:** Writing – review & editing, Investigation. **Militsa Bitzani:** Writing – review & editing, Conceptualization. **Aggelos K. Katsaggelos:** Conceptualization, Supervision, Writing – review & editing. **Nicos Maglaveras:** Writing – original draft, Supervision, Conceptualization, Funding acquisition.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The datasets used and analyzed during the current study are available from: https://figshare.com/s/e5af036d5ca46150eac4.

## Acknowledgments

This work was partially supported by the Horizon 2020 Framework Programme of the European Union project WELMO (grant agreement number 825572), by the FCT - Foundation for Science and Technology, I.P./MCTES through national funds (PIDDAC), under the projects UIDB/04501/2020 and CISUC - UID/CEC/00326/2020, FCT project Lung@ICU under grant reference DSAIPA/AI/0113/2020, FCT Ph.D. scholarships SFRH/BD/135686/2018 and DFA/BD/4927/2020, Fundo Europeu de Desenvolvimento Regional (FEDER) through Programa Operacional Competitividade e Internacionalização (COMPETE) and POCI-01-0145-FEDER-007628—iBiMED.

## Appendix

See Figs. 5–8.

## References

Aboutalebi, H., Pavlova, M., Shafiee, M. J., Sabri, A., Alaref, A., & Wong, A. (2021). Covid-net cxr-s: Deep convolutional neural network for severity assessment of covid-19 cases from chest x-ray images. *Diagnostics*, *12*(1), 25.

Akrami, H., Joshi, A. A., Li, J., Aydore, S., & Leahy, R. M. (2020). Brain lesion detection using a robust variational autoencoder and transfer learning. In *2020 IEEE 17th international symposium on biomedical imaging (ISBI)* (pp. 786–790). IEEE.

Aljouie, A. F., Almazroa, A., Bokhari, Y., Alawad, M., Mahmoud, E., Alawad, E., et al. (2021). Early prediction of COVID-19 ventilation requirement and mortality from routinely collected baseline chest radiographs, laboratory, and clinical data with machine learning. *Journal of Multidisciplinary Healthcare*, *14*(July), 2017–2033. http://dx.doi.org/10.2147/JMDH.S322431.

Alser, O., Mokhtari, A., Naar, L., Langeveld, K., Breen, K. A., El Moheb, M., et al. (2021). Multisystem outcomes and predictors of mortality in critically ill patients with COVID-19: demographics and disease acuity matter more than comorbidities or treatment modalities. *Journal of Trauma and Acute Care Surgery*, *90*(5), 880–890.

Armstrong, R., Kane, A., Kursumovic, E., Oglesby, F., & Cook, T. M. (2021). Mortality in patients admitted to intensive care with COVID-19: an updated systematic review and meta-analysis of observational studies. *Anaesthesia*, *76*(4), 537–548.

Bae, J., Kapse, S., Singh, G., Gattu, R., Ali, S., Shah, N., et al. (2021). Predicting mechanical ventilation and mortality in COVID-19 using radiomics and deep learning on chest radiographs: A multi-institutional study. *Diagnostics*, *11*(10), 1812.

Cheng, J., Sollee, J., Hsieh, C., Yue, H., Vandal, N., Shanahan, J., et al. (2022). COVID-19 mortality prediction in the intensive care unit with deep learning based on longitudinal chest X-rays and clinical data. *European Radiology*, 1–11.

Clevert, D.-A., Unterthiner, T., & Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289.

Cohen, J. P., Dao, L., Roth, K., Morrison, P., Bengio, Y., Abbasi, A. F., et al. (2020). Predicting covid-19 pneumonia severity on chest x-ray with deep learning. *Cureus*, *12*(7).

Ferrando, C., Mellado-Artigas, R., Gea, A., Arruti, E., Aldecoa, C., Bordell, A., et al. (2020). Patient characteristics, clinical course and factors associated to icu mortality in critically ill patients infected with SARS-CoV-2 in Spain: a prospective, cohort, multicentre study. *Revista Española de Anestesiología y Reanimación (English Edition)*, *67*(8), 425–437.

Gallo Marin, B., Aghagoli, G., Lavine, K., Yang, L., Siff, E. J., Chiang, S. S., et al. (2021). Predictors of COVID-19 severity: a literature review. *Reviews in Medical Virology*, *31*(1), 1–10.

Gourdeau, D., Potvin, O., Biem, J. H., Cloutier, F., Abrougui, L., Archambault, P., et al. (2022). Deep learning of chest X-rays can predict mechanical ventilation outcome in ICU-admitted COVID-19 patients. *Scientific Reports*, *12*(1), 1–10. http://dx.doi.org/10.1038/s41598-022-10136-9.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2021). Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Vol. 2017-January* (pp. 2261–2269). IEEE, http://dx.doi.org/10.1109/CVPR.2017.243, arXiv:1608.06993.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Snin, H. H., Zheng, Q., et al. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *454*(1971), 903–995. http://dx.doi.org/10.1098/rspa.1998.0193.

Jaeger, S., Candemir, S., Antani, S., Wáng, Y.-X. J., Lu, P.-X., & Thoma, G. (2014). Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *American Journal of Roentgenology*, *4*, 475.

Kilintzis, V., Beredimas, N., Kaimakamis, E., Stefanopoulos, L., Chatzis, E., Jahaj, E., et al. (2022). CoCross: An ICT platform enabling monitoring recording and fusion of clinical information chest sounds and imaging of COVID-19 ICU patients. In *Healthcare, Vol. 10* (p. 276). MDPI.

Kilintzis, V., Chouvarda, I., Beredimas, N., Natsiavas, P., & Maglaveras, N. (2019). Supporting integrated care with a flexible data management framework built upon linked data, HL7 FHIR and ontologies. *Journal of biomedical informatics*, *94*, Article 103179.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Kwon, Y. J. F., Toussie, D., Finkelstein, M., Cedillo, M. A., Maron, S. Z., Manna, S., et al. (2021). Combining initial radiographs and clinical variables improves deep learning prognostication in patients with COVID-19 from the emergency department. *Radiology: Artificial Intelligence*, *3*(2), Article e200098. http://dx.doi.org/10.1148/ryai.2020200098.

Lartillot, O., & Toiviainen, P. (2007). Mir in matlab (II): A toolbox for musical feature extraction from audio. *Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007*, 127–130.

Leoni, M. L. G., Lombardelli, L., Colombi, D., Bignami, E. G., Pergolotti, B., Repetti, F., et al. (2021). Prediction of 28-day mortality in critically ill patients with COVID-19: Development and internal validation of a clinical prediction model. *PLoS One*, *16*(7), Article e0254550.

Lerch, A. (2012). *An introduction to audio content analysis: applications in signal processing and music informatics* (pp. 1–248). Wiley-IEEE Press, http://dx.doi.org/10.1002/9781118393550.

Narin, A., Kaya, C., & Pamuk, Z. (2021). Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural net- works. *Pattern Analysis and Applications*, *24*, 1207–1220.

Noda, A., Saraya, T., Morita, K., Saito, M., Shimasaki, T., Kurai, D., et al. (2020). Evidence of the sequential changes of lung sounds in covid-19 pneumonia using a novel wireless stethoscope with the telemedicine system. *Internal Medicine*, *59*(24), 3213–3216.

Otsu, N. (1979). A threshold selection method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, *C*(1), 62–66.

Pancaldi, F., Pezzuto, G. S., Cassone, G., Morelli, M., Manfredi, A., D'Arienzo, M., et al. (2022). VECTOR: An algorithm for the detection of COVID-19 pneumonia from velcro-like lung sounds. *Computers in biology and medicine*, Article 105220.

Peeters, G. (2004). *A large set of audio features for sound description (similarity and classification) in the CUIDADO project*: *Technical report*, IRCAM, URL: http://www.citeulike.org/group/1854/article/1562527.

Putman, R. K., Gudmundsson, G., Axelsson, G. T., Hida, T., Honda, O., Araki, T., et al. (2019). Imaging patterns are associated with interstitial lung abnormality progression and mortality. *American Journal of Respiratory and Critical Care Medicine*, *200*, 175–183.

Rocha, B. M., Pessoa, D., Cheimariotis, G. A., Kaimakamis, E., Kotoulas, S. C., Tzimou, M., et al. (2021). Detection of squawks in respiratory sounds of mechanically ventilated COVID-19 patients. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, Vol. December* (pp. 512–516). http://dx.doi.org/10.1109/EMBC46164.2021.9630734.

Rocha, B. M., Pessoa, D., Marques, A., Carvalho, P., & Paiva, R. P. (2021). Automatic classification of adventitious respiratory sounds: A (un)solved problem? *Sensors (Switzerland)*, *21*(1), 1–19. http://dx.doi.org/10.3390/s21010057.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 234–241). Springer.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*(C), 53–65. http://dx.doi.org/10.1016/0377-0427(87)90125-7.

Sait, U., KV, G. L., Shivakumar, S., Kumar, T., Bhaumik, R., Prajapati, S., et al. (2021). A deep-learning based multimodal system for Covid-19 diagnosis using breathing sounds and chest X-ray images. *Applied Soft Computing*, *109*, Article 107522.

Serafim, R. B., Póvoa, P., Souza-Dantas, V., Kalil, A. C., & Salluh, J. I. (2021). Clinical course and outcomes of critically ill patients with COVID-19 infection: a systematic review. *Clinical Microbiology and Infection*, *27*(1), 47–54.

Shamout, F. E., Shen, Y., Wu, N., Kaku, A., Park, J., Makino, T., et al. (2021). An artificial intelligence system for predicting the deterioration of COVID-19 patients in the emergency department. *npj Digital Medicine*, *4*(1), http://dx.doi.org/10.1038/s41746-021-00453-0, arXiv:2008.01774.

Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Ko-matsu, K.-i., et al. (2000). Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology*, *174*, 71–74.

Sriram, A., Muckley, M., Sinha, K., Shamout, F., Pineau, J., Geras, K. J., et al. (2021). COVID-19 deterioration prediction via self-supervised representation learning and multi-image prediction. *ArXiv*, 1–17, arXiv:arXiv:2101.04909v2.

Tabik, S., Gómez-Ríos, A., Martín-Rodríguez, J. L., Sevillano-García, I., Rey-Area, M., Charte, D., et al. (2020). COVIDGR dataset and COVID-SDNet methodology for predicting COVID-19 based on chest X-ray images. *IEEE Journal of Biomedical and Health Informatics*, *24*(12), 3595–3605.

Velavan, T. P., & Meyer, C. G. (2020). The COVID-19 epidemic. *Tropical Medicine & International Health*, *25*(3), 278.

Wang, B., Liu, Y., Wang, Y., Yin, W., Liu, T., Liu, D., et al. (2020). Characteristics of pulmonary auscultation in patients with 2019 novel coronavirus in China. *Respiration*, *99*(9), 755–763.

Wehbe, R. M., Sheng, J., Dutta, S., Chai, S., Dravid, A., Barutcu, S., et al. (2021). Deepcovid-XR: an artificial intelligence algorithm to detect COVID-19 on chest radiographs trained and tested on a large US clinical data set. *Radiology*, *299*(1), E167.

Xie, J., Covassin, N., Fan, Z., Singh, P., Gao, W., Li, G., et al. (2020). Association between hypoxemia and mortality in patients with COVID-19. *95*, In *Mayo Clinic Proceedings* (6), (pp. 1138–1147). Elsevier.

Yuki, K., Fujiogi, M., & Koutsogiannaki, S. (2020). COVID-19 pathophysiology: A review. *Clinical Immunology*, *215*, Article 108427.