AUTOMATIC LUNG AERATION ASSESSMENT FOR PEDIATRIC LUNG ULTRASOUND IMAGING

T. LOGENDRAN

TECHNICAL MEDICINE MASTER'S THESIS

MEDICAL IMAGING AND INTERVENTIONS

EXAMINATION COMMITTEE

Chairman
Prof.dr.ir. C.H. Slump

Medical supervisors
drs. S.G.J. Heisterkamp
dr. A.M.A. Willems

Technical supervisor
dr. C.O. Tan

Professional behavior supervisor
drs. N.S. Cramer Bornemann

External member
ir. E.I.S. Hofmeijer

L U M C Leiden University Medical Center

UNIVERSITY OF TWENTE.

AUTOMATIC LUNG AERATION ASSESSMENT FOR PEDIATRIC LUNG ULTRASOUND IMAGING

T. Logendran

25-10-2022

Technical Medicine – Medical Imaging and Interventions

Master's thesis

Examination committee

Chairman

Prof.dr.ir. C.H. Slump

Robotics and Mechatronics, University of Twente

Medical supervisors

drs. S.G.J. Heisterkamp

Pediatric intensivist, Pediatric intensive care unit, Leiden University Medical Center

dr. A.M.A. Willems

Pediatric intensivist, Pediatric intensive care unit, Leiden University Medical Center

Technical supervisor

dr. C.O. Tan

Associate professor, Faculty of Electrical Engineering, Mathematics and Computer Science, Robotics and Mechatronics, University of Twente

Professional behavior supervisor

drs. N.S. Cramer Bornemann

Clinical internships and professional behavior, University of Twente

External member

ir. E.I.S. Hofmeijer

PhD candidate, Faculty of Electrical Engineering, Mathematics and Computer Science, Robotics and Mechatronics, University of Twente

GLOSSARY

B-mode	Brightness mode
cGAN	Conditional generative adversarial network
CNN	Convolutional neural network
CXR	Chest X-ray
DL	Deep learning
GAN	Generative adversarial network
ICC	Intraclass correlation coefficient
ICU	Intensive care unit
LSTM	Long short-term memory
LUS	Lung ultrasound/ultrasonography
M-mode	Motion mode
NICU	Neonatal intensive care unit
PedLUS	Pediatric lung ultrasonography
PICU	Pediatric intensive care unit
PLAPS	Posterolateral alveolar and/or pleural syndrome
RCNN	Recurrent convolutional neural network
RNN	Recurrent neural network
TL	Transfer learning
US	Ultrasound/ultrasonography

PREFACE

In this master's thesis, I present the work that I have done during my graduation internship at the pediatric intensive care unit of the Leiden University Medical Center. To facilitate the adoption of lung ultrasonography for pediatric intensive care, I have developed deep learning software aimed at supporting the observer in the assessment of the images. First, the topic will be introduced, followed by backgrounds on lung pathophysiology, ultrasonography and deep learning. In the following chapters my research methods and results will be presented and this thesis will be concluded with an evaluation on the study and future perspectives. I hope you will enjoy reading my thesis!

Tharanghi Logendran

October 2022

CONTENTS

Glossary **1**

Preface **2**

1 Introduction **5**

2 Lung ultrasonography **7**
2.1 Healthy lungs . . . 7
2.2 Interstitial disease, consolidation and atelectasis . . . 8
2.2.1 Lung ultrasound score . . . 8
2.3 Pleural effusion . . . 11
2.4 Pneumothorax . . . 11

3 Deep learning **13**
3.1 Training strategies . . . 13
3.2 Network architectures . . . 14
3.2.1 Convolutional neural network . . . 14
3.2.2 Generative adversarial network . . . 14
3.2.3 Recurrent neural network . . . 15

4 Materials and Methods **17**
4.1 Inclusion and exclusion criteria . . . 17
4.2 Data acquisition and annotation . . . 17
4.3 Video-based score prediction . . . 18
4.3.1 Problem definition . . . 18
4.3.2 Model definition . . . 19
4.3.3 Loss definition . . . 19
4.3.4 Training strategy . . . 19
4.4 Segmentation mask-based score prediction . . . 19
4.5 Frame-based segmentation . . . 20
4.5.1 Problem definition . . . 20
4.5.2 Model definition . . . 20
4.5.3 Loss definition . . . 22
4.5.4 Training strategy . . . 22
4.6 Performance evaluation . . . 22

5 Results **24**
5.1 Video-based score prediction . . . 24
5.2 Segmentation mask-based score prediction . . . 25
5.3 Frame-based segmentation . . . 27

6 Discussion **29**

7 Aknowledgements **34**

References **35**

1 INTRODUCTION

Critically ill patients in intensive care units (ICU) often suffer from a wide variety of diseases and complications. These include patients in the pediatric ICU (PICU), who therefore are monitored and examined per system of the human body and are treated accordingly. Some of the complications of PICU patients affect the respiratory tract: the lungs and tracheobronchial airways may be severely affected by respiratory infections - either acquired before or during hospitalization - but also by fluid imbalance, e.g. with cardiac and kidney failure causing edema[1]. Several studies have even demonstrated a correlation between fluid overload and increased risk of mortality in critically ill patients[2][3][4]. Therefore monitoring the lungs and pulmonary function are central components in the PICU.

Diseases of the lungs and chest are frequently diagnosed and monitored with chest X-ray (CXR) imaging as a first-line imaging method[5]. However, during CXR imaging, patients are exposed to ionizing radiation; in a single CXR examination, adults receive an effective dose of approximately 0.2 mSv. To put this in context, the International Commission on Radiological Protection recommends a maximum annual dose of 1 mSV per year for the general public[6]. Irradiation is even more harmful for pediatric patients than patients of other age groups, due to their small size and highly radiosensitive tissues and organs, hence increasing the risk of developing malignancies in these tissues[7][8]. A study showed that the risk of developing cancer in infants receiving a single small dose of radiation is two to three times higher than for the average population[9]. It is evident that there is a need for a safer, yet reliable lung and chest imaging method for the PICU population.

An alternative imaging method for CXR is ultrasonography (US). Lung ultrasonography (LUS) has become a widespread diagnostic tool in PICU's and neonatal ICU's (NICU). A study demonstrated that since the introduction of LUS for respiratory distress diagnosis in a NICU, the number of CXR examinations decreased with 30% and the radiation dose decreased from 5.54 µGy to 4.47 µGy per infant over a period of 18 months[5]. Moreover, LUS is a quick, easy to operate, real-time and portable imaging method that can be used as a bedside diagnostic tool. Nonetheless, one of the major challenges of LUS is that there remains high interobserver variability[10]. Additionally, sufficient skill in interpreting LUS images and understanding the clinical implications of the acquired image is attained by adequate practice. A study has shown that residents and senior physicians without any experience in LUS acquired the required skills in making the LUS measurements after 25 supervised measurements[11]. The described required skills consist of recognizing normal aeration, interstitial syndrome, lung consolidation and alveolar edema. Difficulties in the interpretation of LUS images remain a major limiting factor in adoption of the imaging method by medical professionals, opposed to the simplicity and costs of the examinations.

This problem may be overcome with applications in deep learning (DL) for assessment of LUS images. Multiple studies have shown promising results of DL algorithms for LUS image classification and detection of LUS image characteristics[12][13][14][15][16][17][18][19]. Because research in this field has accelerated in the light of follow-up on COVID-19 infection, these mod-

els have only been trained and validated with adult LUS data. There are rather few studies on the pathology of the lungs in the pediatric population and pediatric LUS (PedLUS), let alone on DL algorithms to interpret PedLUS images. It is expected that neural networks similar to those developed for adult LUS can be trained to interpret PedLUS images to support the assessment of pediatric physicians. Therefore the goal of this study was to explore and apply applications of DL for LUS assessment for the pediatric population.

2 LUNG ULTRASONOGRAPHY

Ultrasonography (US) is an imaging modality based on the registration of reflections of emitted sound pulses[20]. These reflections occur at transitions between media with different impedances, and the intensity of those reflections depend on these impedance differences; the impedance difference is higher for high transitions (hyperechogenicity) and lower for low transitions (hypoechogenicity). For medical US, commonly used methods are brightness mode (B-mode) and motion mode (M-mode). In B-mode the strength of the reflection after pulse emission is displayed as a function of propagation time, used to calculate the distance from the reflecting structure to the transducer. A collection of reflections with different intensities results in a grayscale image of the scanned area. M-mode is similar to B-mode, except for that the reflections in just one direction are displayed as the brightness modulation as a function of time. As the reflector moves in M-mode, so will the image. The combination of imaging in B-mode and M-mode during an examination can give detailed anatomical, physiological and functional information about tissues and structures.
Traditionally, ultrasonography has been considered an unsuitable imaging modality to examine the thoracic cage, due to the high prevalence of artifacts caused by the ribs, pleurae and aerated lungs[8][21]. However, as the interpretation of these artifacts could be correlated to specific disease patterns, LUS proved to be a useful diagnostic tool. Especially in pediatrics, the relatively unossified thorax of children and the thin subcutaneous tissue form suitable acoustic windows for PedLUS[8][22].

2.1 Healthy lungs

The first step in examining the lungs with US is the identification of the pleural line, which appears as a hyperechoic subcostal thin line. The image up to the pleural line is a real image, showing the tissues of the skin, subcutaneous tissue and muscles. Starting from the pleural line, the aerated lungs will show artifacts caused by scattering of the ultrasonic waves in air[23]. The pleural line in between two adjacent ribs can be observed as the so-called bat sign as in Figure 2.1[22]. Moreover, reverberation between the ultrasound probe and strongly reflecting pleural line will cause the formation of A-lines, which are parallel projections of the pleural line[15][22]. In well aerated lungs these constant projections are prominently represented, as the lung parenchyma itself will be hypoechoic. Furthermore, a physiological sign called lung sliding can be observed, due to the movement of the parietal pleura against the visceral pleura during inspiration and expiration. This can most evidently be observed when imaging in M-mode, where lung sliding is represented as the seashore sign as in Figure 2.5a[8]. The pleural line, accompanying A-lines and lung sliding indicate healthy, well aerated lung tissue in the examined region.

Figure 2.1: B-mode LUS image made with a linear probe. The pleural line in between two adjacent ribs on LUS causes the bat sign[24].

2.2 Interstitial disease, consolidation and atelectasis

The interstitium of the lungs can volumetrically expand due to several causes, such as pulmonary edema, interstitial lung disease, congestive heart disease and respiratory infections [16][25][26]. Both increased interstitial fluid and liquid accumulation in the alveoli will cause the formation of B-lines on LUS, which are vertical artifacts most likely caused by the isolation of a fluid structure in a tetrahedron of air bubbles[27]. The fluid resonates in response to the ultrasonic pulse, which is detected by the probe and represented as a hyperechoic line[26]. These lines will erase other image characteristics through their path, including A-lines and reach from the pleural line to the lower edge of the image window, moving with respiration. When a large number of B-lines occur, eventually they will appear as compact hyperechoic structures, indicating a near consolidated lung and in further stages, hepatization of the lungs will be observed[21][22]. In hepatized lungs, the alveoli have been collapsed or are filled with fluids to the extent that the lung parenchyma will appear as dense tissue, similar to the ultrasonic representation of liver tissue. Characteristics for extended lung consolidation are bronchograms with shred signs; distinctive irregular patterns in the scanned region[28]. The appearance and severity of the described LUS characteristics can be correlated to the type and severity of different lung pathologies and, together with the clinical presentation of the patient, can provide valuable information about the condition of the lung tissue.

2.2.1 Lung ultrasound score

The LUS score is a semi-quantitative scoring system for the regional aeration of the lungs[11]. The scoring system is practical for monitoring both intra- and interpatient regional lung aeration

and may serve as a quantitative measure when analyzing lung diseases, mechanical ventilation and weaning. In order to assess the lungs thoroughly, the thorax is divided into twelve regions; six regions for each hemithorax as in Figure 2.2. Each region is scored from 0 to 3 based on the degree of pulmonary aeration loss observed with LUS, hence the total LUS score can vary from 0 to 36. A LUS score 0 indicates healthy, aerated lung parenchyma with A-lines and up to two B-lines per view, a score 1 moderate loss of aeration with multiple (three or more) well-separated B-lines per view, score 2 severe loss of aeration with coalescent B-lines, and score 3 complete dense consolidation with or without bronchograms[11][29]. Figure 2.3 illustrates examples of LUS images for each LUS score. The assessment of a region is based on the view of that region with the most severe score. Generally, the most severe scores are assigned to posterolateral regions in supine positioned patients due to gravity. Therefore a crucial area to include in LUS examinations is the so-called posterolateral alveolar and/or pleural syndromes (PLAPS) point[30], shown in Figure 2.4.

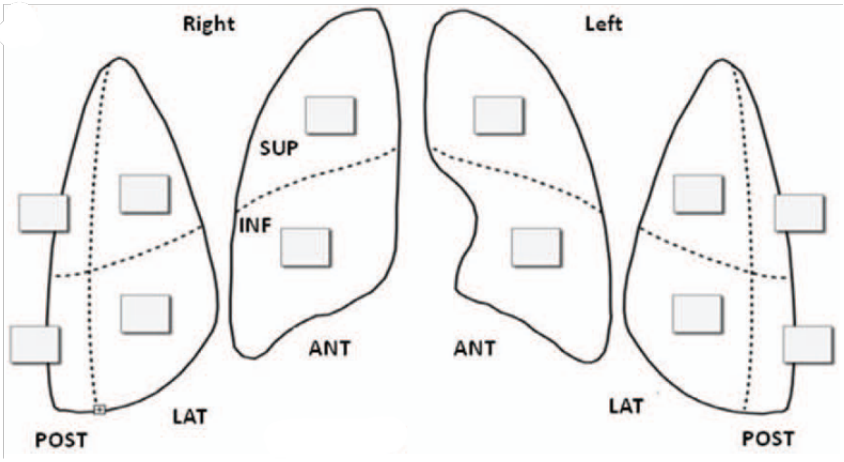


Figure 2.2: Twelve regions of the thorax to be scored on LUS[29]. Each hemithorax is subdivided into six regions. For each explored region, the most severe finding in that region is reported. ANT = anterior; INF = inferior; LAT = lateral; POST = posterior; SUP = superior.

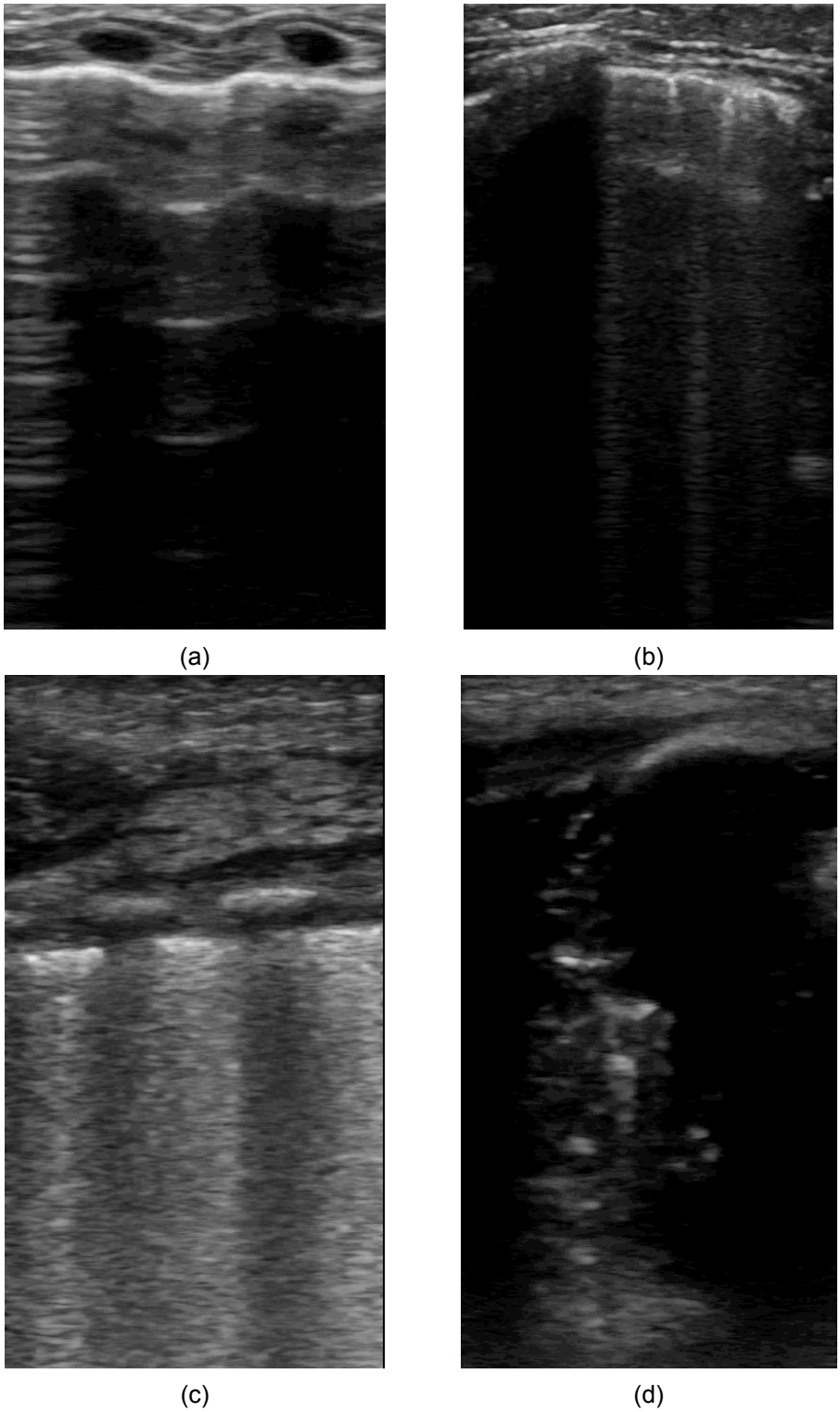


Figure 2.3: Examples of LUS images with (a) score 0, (b) score 1, (c) score 2 and (d) score 3[21]. Image (a) and (b) show examples of A-lines, image (b) containing some B-lines. Image (c) is an example of coalescent B-lines and image (d) depicts an example of the shred sign caused by consolidations.

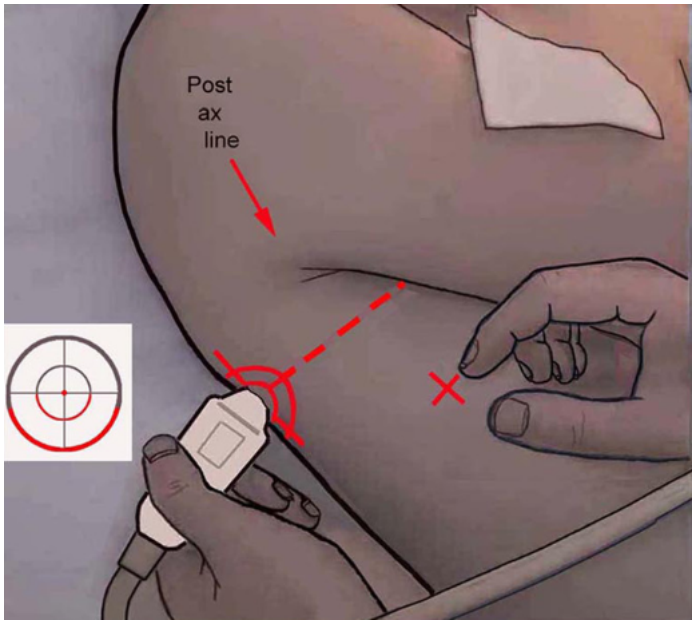


Figure 2.4: Location of the PLAPS point, where the incidence of interstitial fluid, consolidations and pleural effusion is high. Modified from [30].

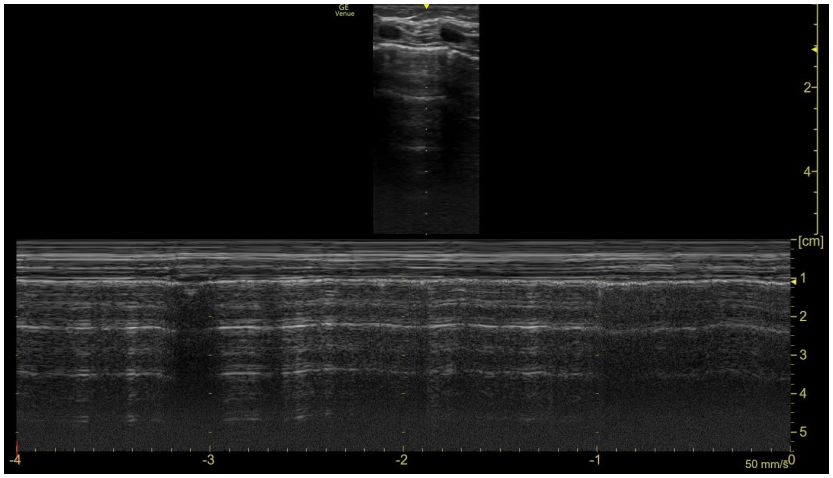
2.3 Pleural effusion

Besides lung aeration assessment, LUS is also a suitable method to evaluate the presence and severity of pleural effusion. Similar to interstitial and intra-alveolar fluids, effusion can best be evaluated at the PLAPS point in supine positioned patients. The appearance of effusion depends on its nature: most transudates and some exudates are anechoic, hence reflections in this anechoic space strongly suggest the presence of exudates. Echogenic effusions are often due to hemothorax or empyema and are sometimes presented with septations within the effusion[31]. The ability to evaluate pleural effusion volumes and to differentiate the consistency of effusions, make LUS a valuable tool in the PICU.

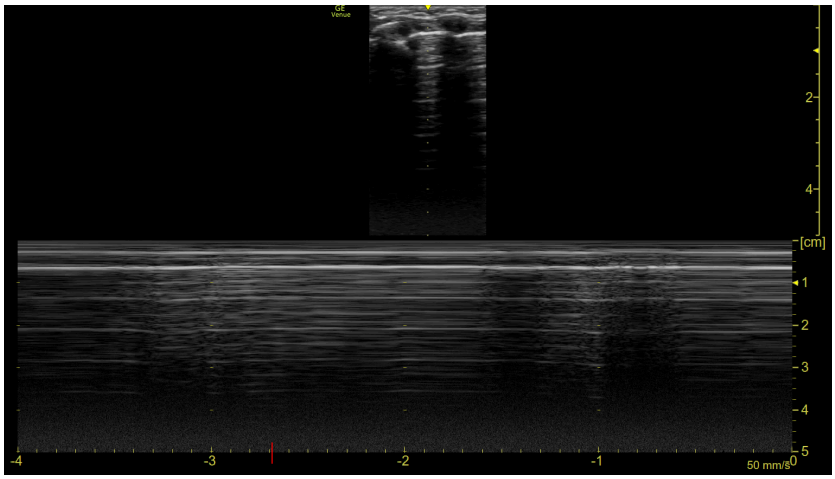
2.4 Pneumothorax

In healthy lungs, the movement of the pleurae can be observed as lung sliding. Absence of lung sliding in the examined area may indicate detachment of the lung from the thoracic cage in the examined area, hence the indication of pneumothorax[21]. The lung point is a specific sign for pneumothorax, indicating the location where the attached part of the lung transitions into the detached part, although the lung point may be missing in severe pneumothorax. Moreover, when imaging in M-mode, the so-called barcode or stratosphere sign as in Figure 2.5b is present instead of the seashore sign, caused by the absence of dynamic lung sliding[22][32]. Furthermore, B-lines are absent, as these are characteristic for pleura-lung boundaries. Finally, no lung pulse can be observed with pneumothorax, which is the rhythmic movement of the pleurae with cardiac oscillations[5]. The lung pulse is usually masked by lung sliding, but can be observed when there is little or no sliding whilst the lung is attached to the thoracic cage[8][33].

There are many features in LUS images that can be measured and analyzed to evaluate the



(a)



(b)

Figure 2.5: M-mode images showing a (a) seashore sign and (b) barcode sign, the latter caused by pneumothorax.

condition and aeration of lung tissue. However, LUS is not a widely used imaging method yet, due to the extensive experience that is required to be able interpret the images. Therefore, to assist the observer in the interpretation of LUS images, the application of DL for automatic LUS assessment has been investigated in this study. In the next chapter, DL will be elaborated upon in more depth, illustrating potential applications for LUS image interpretation.

3 DEEP LEARNING

Deep learning refers to a computer approach to learn and optimize a task based on a certain set of features[34]. The goal of DL is to find an approximation of the underlying correlation of a presented problem, such as a classification, recognition, or segmentation task, based on the provided inputs and desired outputs. When a new input is presented, the model is able to predict an output in the form of the probability for each possible output category[35]. Regarding the assessment of LUS images, there are multiple predictions that can be made from the input, which can either be an image or a video. The output prediction can be a classification task to predict LUS scores, or the detection or segmentation of pathological artifacts. The error between the output probabilities and the desired pattern of probabilities is evaluated with a loss function that is aimed to reduce this error.
Deep neural network training consists of iterations of presenting training data to the model, which consists of inputs and accompanying outputs. During the training process, the model tries to predict outputs and compares these predicted outputs to the desired outputs. These predicted outputs are evaluated with the loss function, after which internal weights are adjusted in order to minimize the loss[36]. The extent to which weights are adjusted is determined by the learning rate[37]. An optimal learning rate must be found, as a low learning rate may take the algorithm too long for optimization and a high learning rate may result in a sub-optimal set of weights or an unstable training process. The learning rate is often decayed during the training process with an optimizer to avoid these challenges[38].
One iteration of presenting all training data to the model is called an epoch[38]. After a set amount of epochs, the network is tested with new, unseen test data and the ability to classify the input data can be evaluated. The intention is for the model to 'learn' patterns that are generalizable, meaning it will perform well on unseen test data. When this fails, overfitting occurs; a common problem in the training process[38]. This problem tends to occur when the number of free parameters, e.g. when the amount of weight connections, is too large compared to the size of the training data, as it memorized nuances of the data but fails to recognize significant patterns for unseen test data[38]. Overfitting can be prevented by limiting the amount of epochs in the training stage until the variation of error becomes sufficiently small[39]. Another method is to enlarge the data set, but due to limitations in data collection this is often not realizable. A commonly used solution for this problem is data augmentation, which is the application of transformations to the available data set. In the case of image data, this can include rotations, zooming, shifting, flipping and applying color space transformations to the images[40][41][42]. Applying these transformations in some cases may prevent overfitting and improve the test accuracy[43].

3.1 Training strategies

DL network training can either be supervised, unsupervised or semi-supervised[44]. In supervised learning all training data is labeled, meaning that for each training input a desired output

is determined. During unsupervised learning, the model learns important features to discover undefined, unknown relationships or structures within the input data. Semi-supervised learning is a combination of the two where the network is trained with partially labeled data. One of the major advantages of semi-supervised and unsupervised learning is that there is little or no need for the time consuming process of labeling data. However, unsupervised learning is a very complex process requiring the network to sort data without manual intervention[45]. In the context of medical imaging and LUS image assessment, an advantage of supervised learning is that observers can set clear class or segmentation boundaries, based on medically defined classes and definitions, rather than technical definitions. With unsupervised learning the model is trained to find patterns and connections in the data by itself, which is not preferable in the interpretation of LUS measurements.

3.2 Network architectures

3.2.1 Convolutional neural network

A popular type of deep learning network is the convolutional neural network (CNN), which is designed to process spatially dependent grid-structured inputs making this architecture well suited to process 2D data, such as images. Images have spatial dependencies, as pixel values are strongly correlated to the values of surrounding pixels. A defining characteristic of CNNs is the convolution operation, which is an operation where a filter is shifted over the input values of the image pixels and the dot product of those is computed with each shift[38][46]. These shifts are continued until the whole image is covered and a so-called feature map is computed. CNNs learn to detect different features of the images using tens to hundreds of hidden layers applying convolutions, increasing the complexity of the learned image features with each hidden layer[45]. An important aspect of CNNs is that feature detection is not spatially dependent, because the images are analyzed regionally for different features[47].
There are multiple studies on the use of CNNs for LUS interpretation. Panicker et al.[16] described a model based on a U-net architecture for segmentation of A-lines, reduction in the pleural thickness and single or multiple B-lines to a coalescent appearance. In a comparable study[19] an Inception v2 based CNN was developed to identify B-lines, merged B-lines, lack of lung sliding, consolidations and pleural effusion. They collected data as videos which were used for training as sequential temporal and individual frames. Roy et al.[12] also collected both images and videos that were labeled according to a LUS scoring system. Their feature detection was performed with a U-net baseline model and they have implemented a spatial transformer network to localize pathological artifacts. Besides feature detection, video-level grading was used to predict a LUS score for the entire video. Another similar model has been designed and trained for the detection of B-lines, which was able to classify images as containing and not containing B-lines, then localize B-lines in images where they are present[15]. La Salvia et al.[14] used class activation mapping to highlight the parts that were decisive for the classification task. They used transfer learning (TL) on four different networks to make COVID diagnoses based on LUS images. In TL, the learned features of a pre-trained model on one problem are used for a new, similar problem using a partially related or unrelated dataset to overcome the obstacle of insufficient training data[45].

3.2.2 Generative adversarial network

Another common DL network is the generative adversarial network (GAN), which is designed for image-to-image translations. The model consists of a generator model, to generate new

plausible synthetic images that are similar to the desired outputs, and a discriminator model that is presented both real desired outputs and generated outputs, after which it classifies those images as real or generated. The discriminator is updated directly, while the generator model is updated through the discriminator. The models are therefore trained simultaneously, as the generator tries to better 'fool' the discriminator and the discriminator tries to better identify the fake images[48]. After each iteration, a discriminator loss and generator loss are defined which both should be minimized during training. A schematic representation of a GAN is provided in Figure 3.1.

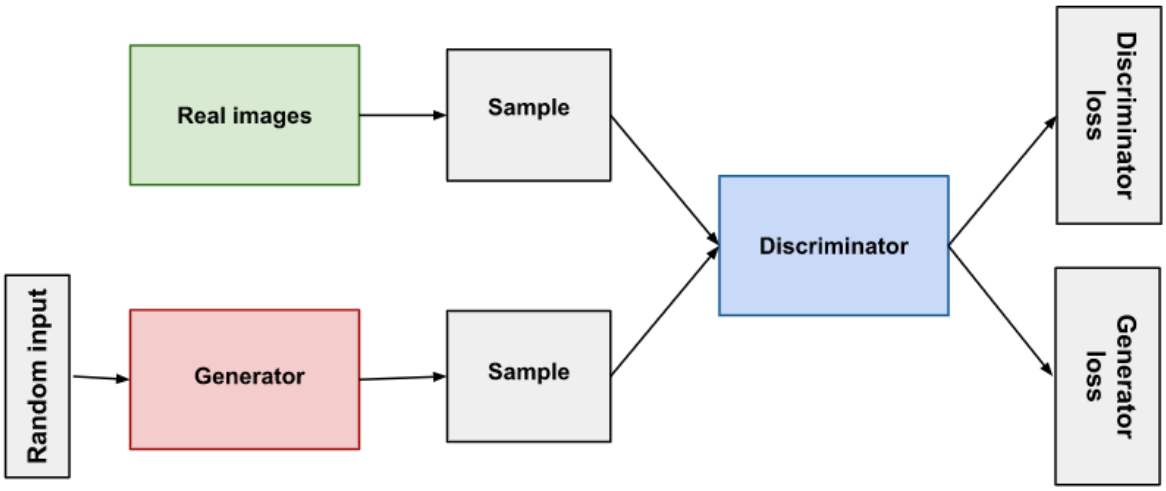


Figure 3.1: Schematic overview of a GAN[49]. The generator attempts to generate an output, similar to the desired output. The discriminator tries to distinguish generated outputs from the actual desired outputs.

There have been few studies on the use of GANs for thoracic image generation. Loey et al. and Ali et al. have proposed a GAN to generate CXR images to increase their database, which they have used to train a CNN to detect COVID-19 related features[50][51]. In another study a GAN was used to generate data to classify LUS images as healthy, bacterial pneumonia, COVID-19 infection or fake[52]. In this case, the generated image is supposed to be classified fake. The output of the generator may also be useful to create outputs containing clinical features of interests, such as B-lines and consolidations. Focusing on this generating aspect of GANs, the model could be trained to generate plausible segmentation masks for LUS images that can assist the physician in the interpretation of the images.

3.2.3 Recurrent neural network

When having to process inputs of variable lengths, recurrent neural networks (RNNs) are a useful type of neural network, where previous outputs are weighed in together with current inputs[53]. RNNs are commonly used for sequential data and in the case of videos, temporal data. They are distinguished by their 'memory' as they use information from prior inputs to influence the current input and output[54]. In the context of LUS, RNNs may be valuable models for video classification, as features from LUS videos contributing to the LUS score are dynamic and are not present in every frame of the video. Therefore the entire sequence of frames should be considered in assigning a LUS score to a video.
To our knowledge there are no studies on LUS classification using RNNs. Two studies describe something similar, where a hybrid structure using a CNN and a long short-term memory (LSTM),

a neural network similar to an RNN, was used[13][55]. First the data was passed through the CNN part of the network, after which it was passed through the LSTM to perform the final predictions based on the joint weight vectors generated from both parts.

4 MATERIALS AND METHODS

In this proof of concept study we aimed to explore whether LUS can become a more accessible technique to use in daily practice at the PICU and to reduce the influence of interobserver variability in LUS assessment. Therefore we have designed two DL models: one for video-based classification into LUS scores and one for frame-based segmentation of clinical features in LUS images. Both the clinical features and LUS score predictions were expected to assist physicians in the interpretation of the LUS measurements in terms of aeration of the lungs.

4.1 Inclusion and exclusion criteria

PedLUS data was collected from 33 PICU patients from the Leiden University Medical Center between January and May 2022. The age of the population at the PICU varies from 0 to 18 years old, which was the age group of the study population. There are no significant contraindications for LUS and patients were not differentiated by underlying pathologies.

4.2 Data acquisition and annotation

A total of 506 videos of 4 seconds each was collected. In previous studies, probe selection was based on patient sizes, which sometimes resulted in either having to train separate networks for each probe or having to apply image modifications due to pluriformity of the data[13][16]. In this study, all data was collected with two different linear probes between 3-20 MHz with a standard mid-range ultrasound scanner (Venue GO, GE Healthcare, Chicago, IL, USA) which were also used based on the size of the patient. As image sizes were the same for both probes, the data has not been distinguished by the used probe.
In order to image the lungs completely, the thorax was divided into twelve different scanning areas as in Figure 2.2. Each region was completely scanned, after which the most severe observed LUS score in that area was assigned to that region. Prior to the study we have concluded that the LUS score definitions lack in the assessment of small consolidations in regions that are generally well aerated, hence we have defined the LUS scores as in Table 4.1 when labeling data for this study. Videos that only showed areas with pneumothorax were excluded, as the tissue underneath cannot be assessed with US. All LUS videos were assigned a LUS score by at least two independent observers until consensus was reached.

Table 4.1: LUS score definitions in this study.

LUS score	B-lines	Consolidations
0	No significant B-lines (0-2 lines)	Non
1	Separate B-lines (< 3 lines)	Small subpleural consolidations (appear as 'pinheads')
2	Coalescent B-lines	Bigger consolidations with < 50% loss of aerated parenchyma
3	-	Loss of > 50% aerated parenchyma

In cases where no LUS score could be assigned due to poor image quality or ambiguity in the image, the video was labeled 'indeterminable' (ID). However, there were little videos that could be classified as ID. As we believe that the ability to identify uninformative videos labeled as ID is important, 20 arm US videos were collected. These videos clearly do not contain LUS features and therefore are uninformative, a method that is proven to be effective in previous work[18]. Figure 4.1 depicts examples of arm US frames. From all collected videos, 800 frames were randomly picked for the detection of clinical features, which were manually semantically labeled using Label Studio data labeling software[56]. All software in this study has been written in Python version 3.10.0 in PyCharm IDE (JetBrains s.r.o., Prague, Czech Republic) and can be found at *https://github.com/TharanghiLogendran/LUS_assessment*.

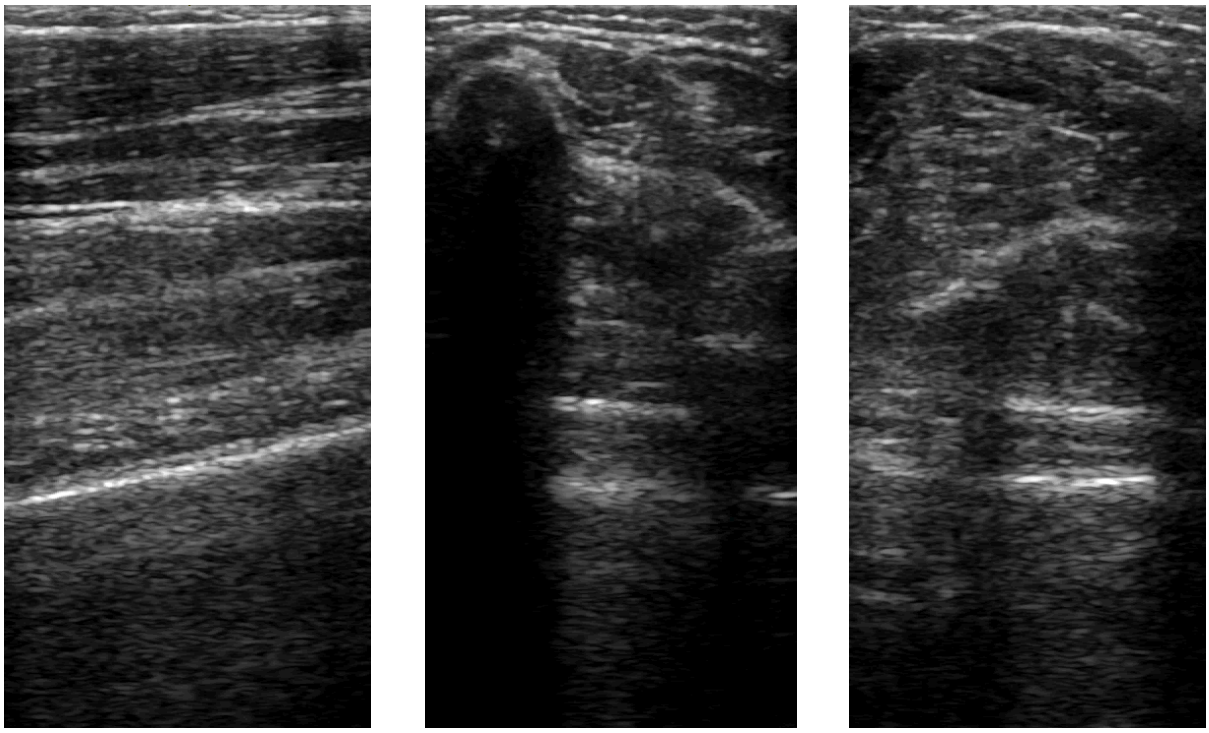


Figure 4.1: Examples of arm US frames used in this study, labeled as 'indeterminable'.

4.3 Video-based score prediction

LUS videos each consist of approximately 180 frames, containing spatial information, and the sequence of those frames contain temporal information. Both aspects are considered useful in classifying the videos into LUS scores, as features from LUS videos contributing to the LUS score are dynamic and are not present in every frame of the video. The proposed model has a hybrid architecture that consists of a CNN for processing spatial information and recurrent layers for processing temporal information, resulting in a recurrent convolutional neural network (RCNN).

4.3.1 Problem definition

Let X be the input space and Y the set of possible scores, so $Y = \{ID, 0, 1, 2, 3\}$. During training, we have a training set $S = \{(x_n, y_n)\}_{n=1}^{N}$ where $x_n \in X$ and $y_n \in Y$. As the input

consists of videos - sequences of frames - we can state that $x_n = (x_1, x_2, ..., x_T)$. For each input video x we want to find the probability $p(y|n)$ for having score y.

4.3.2 Model definition

As the reproducibility of previous work was an elaborate and time consuming process, we have chosen to write all code and train networks from scratch. We are interested in learning and mapping $X \rightarrow Y$. RNNs handle the sequences by having a recurrent hidden state whose activation at each time step is dependent on that of the previous step. Given a sequence x_n, the proposed RCNN updates its recurrent hidden state h_t by

$$h_t = \begin{cases} 0, & t = 0 \\ \phi(h_{t-1}, x_t), & otherwise \end{cases} \tag{4.1}$$

and the probability for score y is defined as

$$p(y|x_1...x_T) = softmax(W_{hy}h_t + b_y) \tag{4.2}$$

with weights W and biases b to be learned.

4.3.3 Loss definition

The sparse categorical loss function is defined as

$$Loss = -\sum_{i=1}^{n} y_i log(p_i) \tag{4.3}$$

where y_i is the true score and p_i is the probability for the i^{th} class for n classes[57]. In sparse categorical cross-entropy, labels are integer encoded, so in this case for the LUS scores $Y = \{1, 2, 3, 4, 5\}$. The aim was to minimize the loss.

4.3.4 Training strategy

From all videos, 75% was used for training, 25% for testing. From each video of approximately 180 frames, 80 frames were extracted to reduce computing time, prevent overfitting and to reassure every sequence was of the same length. The frames were standardized to a fixed image size 224x224 pixels. The frames have been pre-processed by extracting meaningful features from the frames with an Inception v3 base model[58]. The outputs from the CNN were then passed through the RNN. The videos were also augmented to enlarge the data set and to expose the model to different aspects of the training data, by flipping them horizontally and rotating them up to 30°. Each video was augmented four times, which led an increase by a factor 5. The network was trained on a GeForce GTX 1080 Ti 11 GB GPU (nVidia, Santa Clara, CA, USA) and after the training process the best performing model was saved.

4.4 Segmentation mask-based score prediction

RNNs are relatively complex models and to test whether this complexity is required to classify LUS images, we have implemented a simple logistic regression model to perform a similar classification task. Therefore, pathological artifacts, including B-lines, consolidations and atelectasis, have been manually segmented from 464 individual LUS frames that have been randomly

selected from all collected videos. The pixel counts of the created segmentation masks have been correlated to the LUS scores that have been assigned to the individual frames. As the segmentation masks for score 0 and ID would be identical using this method, ID was excluded from the classification categories. The model was trained with 75% of all data, 25% was kept for testing.

4.5 Frame-based segmentation

For the semantic segmentation of clinical features, individual frames have been segmented by a GAN. These segmentation masks included A-lines, B-lines, consolidations and atelectasis. Two networks were trained simultaneously: one that generated images and one that discriminated real images from generated images[59].

4.5.1 Problem definition

Let $x = \mathbf{R}^{i \times j}$ be the input image and y be the output segmentation mask. The image dimensions are here denoted as ixj. In this case, for semantic segmentation we distinguish two different scores assigned to pixels, denoted as $Y = \{0, 1\}^{i \times j}$, as pixels classification is binary: they are either part of the clinical features, so the segmentation mask, or the background. During training we again have a training set $S = \{(x_n, y_n)\}_{n=1}^{N}$ where $x_n \in X$ and $y_n \in Y$.

4.5.2 Model definition

The aim was to learn and map $X \rightarrow Y$. First an initial input x_r^a was provided to the generator G, which is a real input in domain A. This input consists of random noise sampled from a prior distribution $p(z)$ to set initial weights. The generated output of G in domain B is x_g^b and is expected to be similar to the real sample x_r^b that is drawn from the real data distribution $p_r(x)$. We have provided the generator the additional information that the input is an image, so that we expect generated outputs with image properties. This type of GAN is also referred to as conditional GAN (cGAN) which's generation process can be expressed as $x_g = G(z, c)$[59]. An overview of the data flow through the GAN is depicted in Figure 4.2

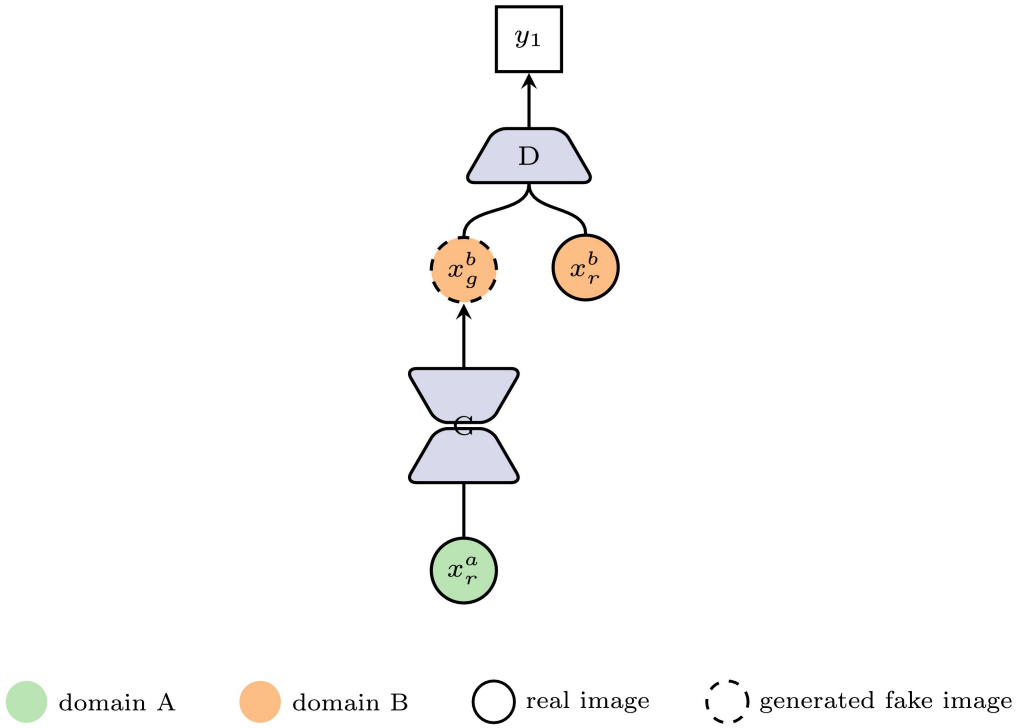


Figure 4.2: Data flow in a GAN. First an initial input is passed through the generator, which simulates an output x_g similar to x_r. The discriminator outputs a probability of the input being real or fake. Modified from [59].

The function learned by G can be denoted as

$$x_g = G(z; \theta_g) \tag{4.4}$$

parameterized by θ_g. Next in the data flow of the GAN, the input to discriminator D is either a real or generated sample. The output of D is y_1 and is a single value indicating the probability of the input being a real or fake sample. The function learned by D, parameterized by θ_d, is denoted as

$$y_1 = D(x; \theta_d) \tag{4.5}$$

The generated samples from G form a distribution $p_g(x)$, which is desired to be an approximation of $p_r(x)$ after successful training. The objective of D is to differentiate x_g and x_r, whereas G is trained to fool D as much as possible. In the training process, information is back propagated from D to G, so G adapts its parameters in order to improve the output images to fool D. The training objectives of D and G can be expressed as

$$\begin{aligned} \mathcal{L}_D^{GAN} &= \max_D \mathbb{E}_{x_r \sim p_r(x)}[log D(x_r)] + \mathbb{E}_{x_g \sim p_g(x)}[log(1 - D(x_g))], \\ \mathcal{L}_G^{GAN} &= \min_G \mathbb{E}_{x_g \sim p_g(x)}[log(1 - D(x_g))]. \end{aligned} \tag{4.6}$$

The desired outcome after training is that samples formed by G approximate the real data distribution $p_r(x)$[59].

4.5.3 Loss definition

The binary cross-entropy loss function is defined as

$$Loss = -\sum_{i=1}^{n} y_i log(p_i) = -[ylog(p) + (1-y)log(1-p)] \tag{4.7}$$

where y_i is the true score and p_i is the probability for the i^{th} class[57]. In this case $Y = \{0, 1\}$, as pixel classification is binary (either part of the segmentation mask or background). The aim was to minimize the loss.

4.5.4 Training strategy

Segmentation masks for 800 randomly selected frames from all collected videos were created. For the frame-based segmentation, frames were re-scaled to an image size of 256x512 pixels. From all data, 80% was used for training and 20% was kept for testing. Performance was evaluated after every 100 training batches on a A40 48GB GPU (nVidia, Santa Clara, CA, USA).
An overview of all options and hyperparameters for training the RCNN and GAN is shown in Table 4.2.

Table 4.2: Options and hyperparameters for model training.

	RCNN classification	GAN segmentation
Initial learning rate	0.001	0.0002
Batch size	64	64
Epochs	120	120
Optimizer	Adam	Adam
Loss function	Sparse categorical cross-entropy	Binary cross-entropy

4.6 Performance evaluation

The performance of the RCNN was evaluated on sensitivity, specificity, accuracy, precision and F1-scores for each output class. These evaluation parameters are determined by true positive (TP), true negative (TN), false positive (FP) and false negative (FN) predictions of the model.

$$Sensitivity = \frac{TP}{TP + FN} \tag{4.8}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4.9}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{4.10}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.11}$$

$$F1 - score = 2 * \frac{Precision * Sensitivity}{Precision + Sensitivity} \tag{4.12}$$

The performance of the GAN was evaluated on Dice similarity coefficients (DSC) and mean of squared errors (MSE). Both metrics are defined by the true image x and the predicted image y.

$$DSC = \frac{2|x \cap y|}{|x| + |y|} \tag{4.13}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y - x)^2 \tag{4.14}$$

5 RESULTS

Table 5.1 provides an overview of demographic variables of the study population.

Table 5.1: Demographic variables of the study population.

Variable	**Number (mean ± SD)**
Females	7
Males	26
Age at first examination (months)	27.1 ± 57.6
LUS examinations per patient	3 ± 3

5.1 Video-based score prediction

Figure 5.1 provides an overview of the total amount of videos used for both training and testing. These include collected and augmented videos.

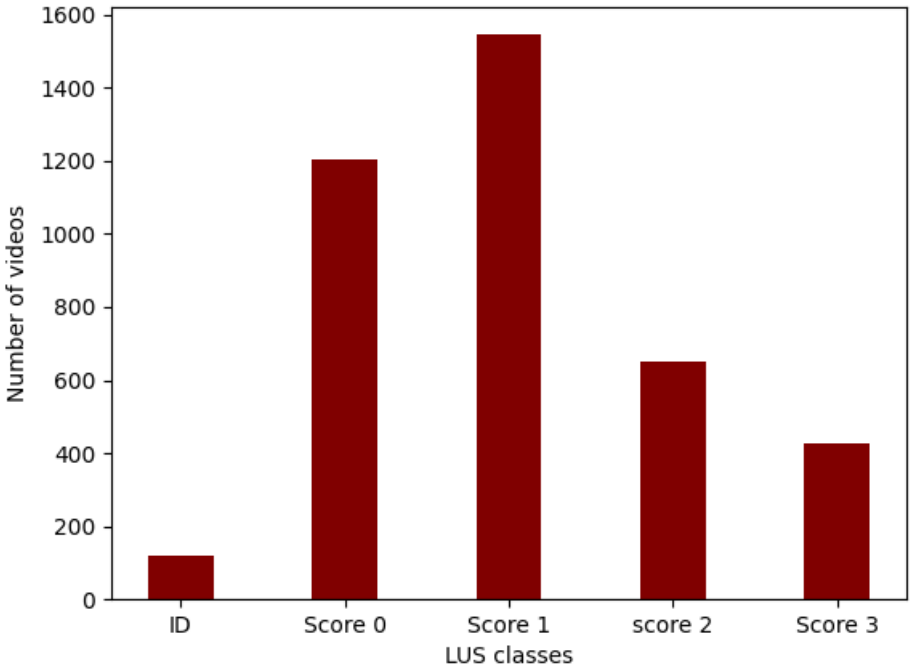


Figure 5.1: Total number of videos for each class, including lower arm ultrasounds and augmented videos, which are randomly flipped or rotated up to 30°.

The results of the LUS video classification by the RCNN are presented in Figure 5.2 and Table 5.2.

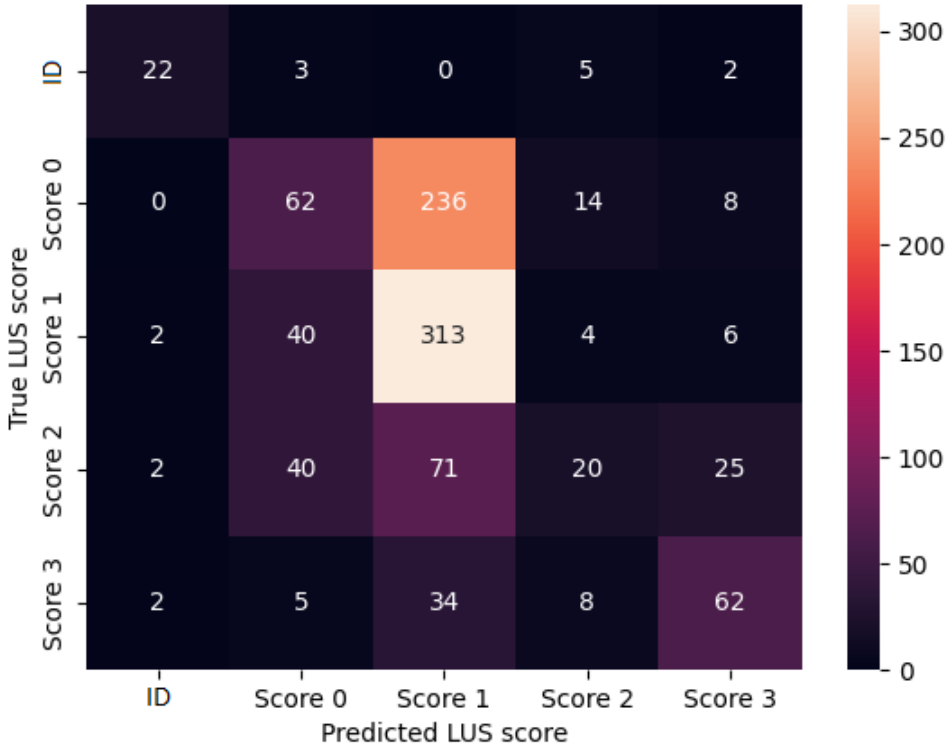


Figure 5.2: Confusion matrix for the classification performance of the RCNN.

Table 5.2: LUS video classification results for each class.

LUS score	Sensitivity (%)	Specificity (%)	Accuracy (%)	Precision (%)	F1-score (%)
Indeterminable	69	99	98	79	73
Score 0	19	87	65	41	26
Score 1	86	45	60	48	61
Score 2	13	96	83	39	19
Score 3	56	95	91	60	58
Average	49	84	79	53	47

5.2 Segmentation mask-based score prediction

The data distribution of the data used for the logistic regression model is presented in Figure 5.3. The results of the classification using this model are presented in Figure 5.4 and Table 5.3.

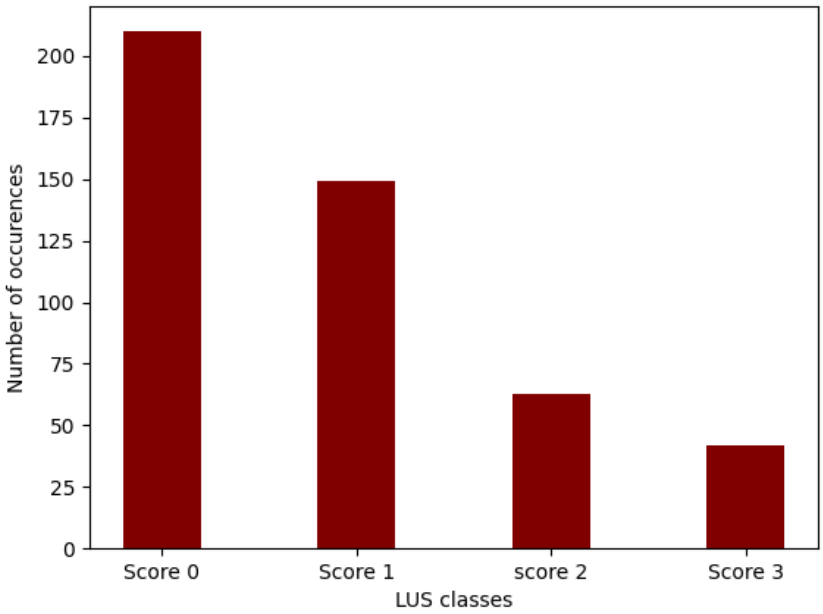


Figure 5.3: Incidence of each LUS score in the data containing segmentation masks used for the logistic regression model.

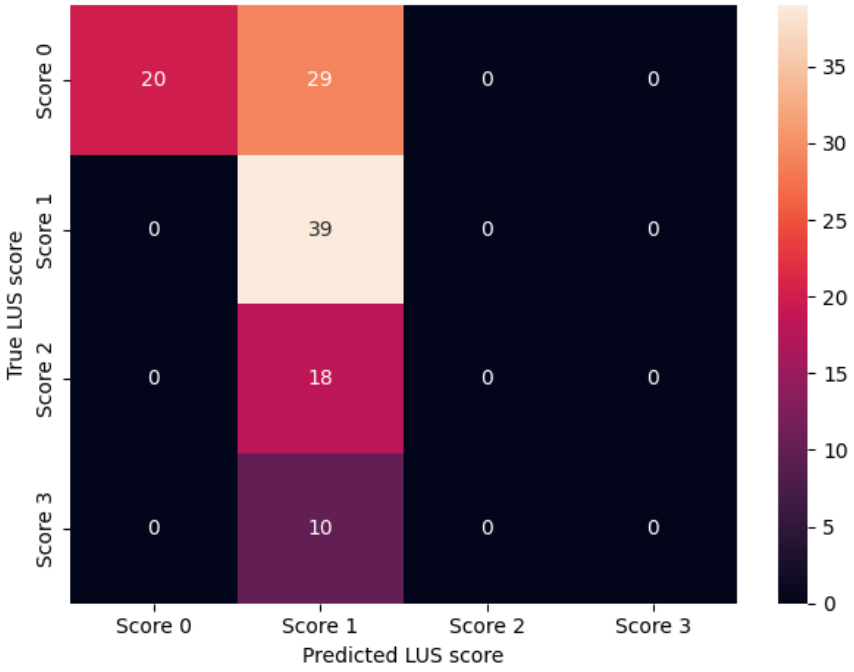


Figure 5.4: Confusion matrix for the classification performance with logistic regression.

Table 5.3: LUS video classification results for each class using a logistic regression model.

LUS score	Sensitivity (%)	Specificity (%)	Accuracy (%)	Precision (%)	F1-score (%)
Score 0	41	100	75	100	58
Score 1	100	26	51	41	58
Score 2	0	100	84	-	-
Score 3	0	100	91	-	-
Average	35	82	75	-	-

5.3 Frame-based segmentation

Figure 5.5 shows the generator loss after each training batch. Table 5.4 shows the segmentation performance of the two best performing models. Examples of LUS frames, the desired segmentation masks and segmentation masks generated by the best performing model are depicted in Figure 5.6.

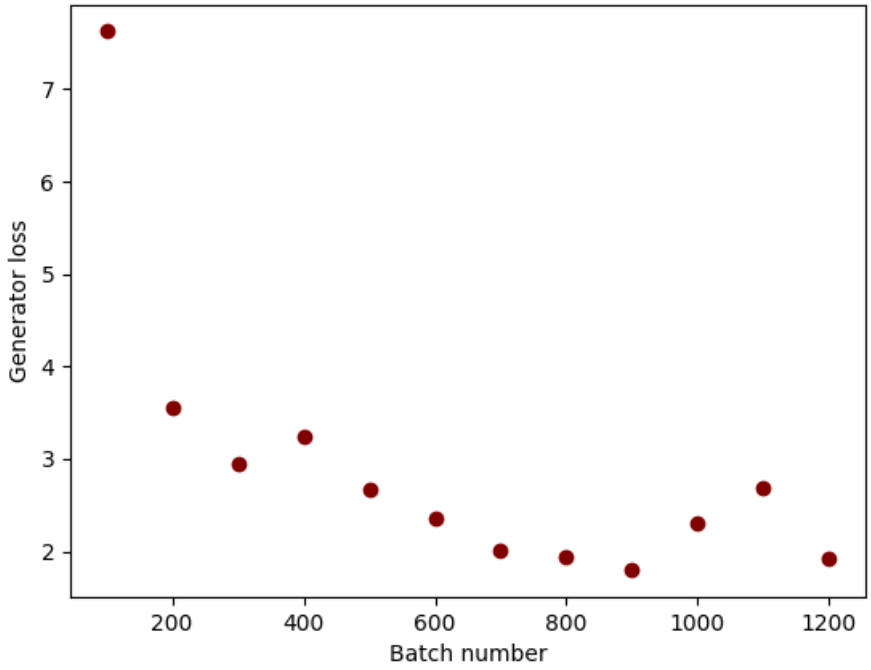
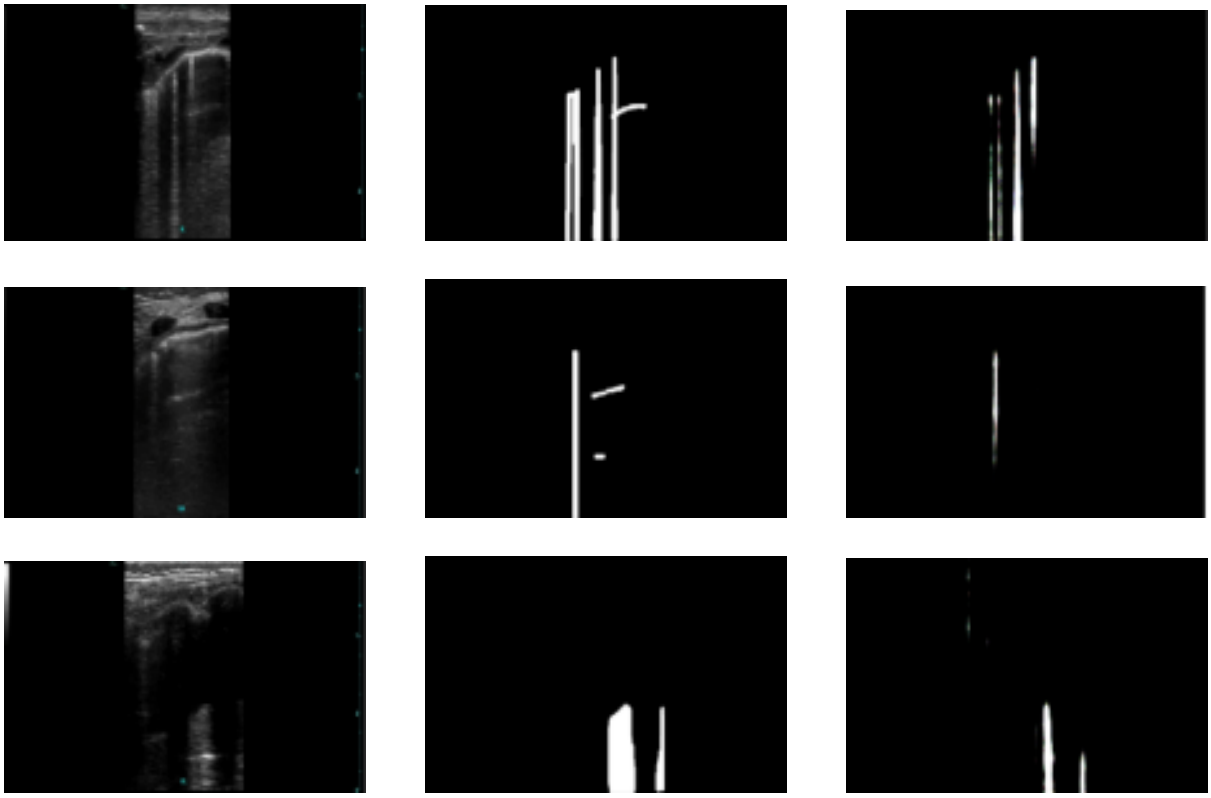


Figure 5.5: Generator loss as a function of the batch number. After every 100 batches the model was saved.

Table 5.4: LUS frame-based segmentation performance.

Saved model	DSC (mean $\pm$ SD)	MSE (mean $\pm$ SD)	Generator loss
model_000900	0.9695 $\pm$ 0.0291	0.0251 $\pm$ 0.0253	1.795
model_001200	0.9632 $\pm$ 0.0291	0.0303 $\pm$ 0.0261	1.932



(a) Individual LUS frames.

(b) Expected segmentation masks.

(c) Generated segmentation masks.

Figure 5.6: Examples of (a) individual LUS frames, (b) the desired segmentation masks and (c) segmentation masks generated by the GAN (model_000900).

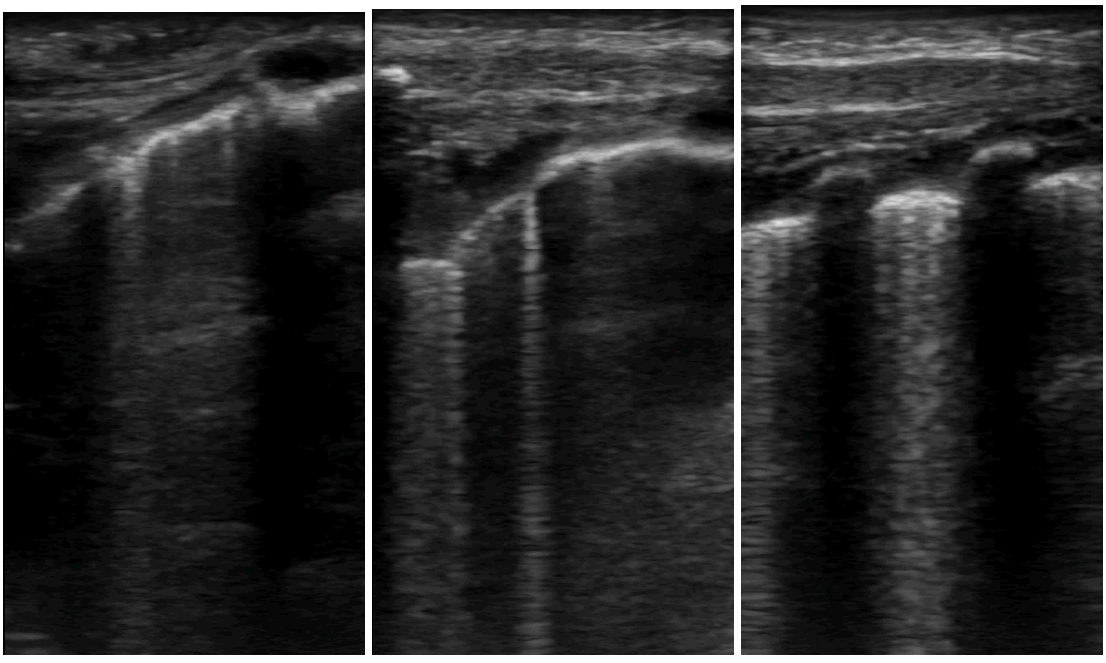
6 DISCUSSION

In this study we proposed a RCNN that classifies PedLUS videos into LUS scores and a GAN that segments A-lines, B-lines, consolidation and atelectasis from individual PedLUS frames. The predictions made by the models are aimed to assist physicians in their interpretation of LUS images to give them an impression of the regional aeration of the lungs. The aim was to facilitate the substitution of chest X-ray imaging by lung ultrasonography in pediatric intensive care.

The classification results in Table 5.2 and Figure 5.2 show that the model classifies most videos as having score 1. This resulted in a high sensitivity for that score: many videos labeled with score 1 were also predicted to have score 1, but the specificity was rather low, hence the model was not able to properly indicate when a video did not have a score 1. For scores 0, 2 and 3 it is the opposite; the model is able to predict when the video does not have one of those scores, but it has more difficulties in making the right prediction when the video does have one of those scores (ergo the sensitivity is low). For scores 0 and 2 the sensitivity (19% and 13% respectively) is considered insufficient for clinical use. Remarkably, for indeterminable videos both the sensitivity and specificity are relatively high. In fact, all outcome measures are high for ID; the results imply that the model is best at predicting when the class is not ID and good at predicting when it is ID. An explanation for these high results compared to the other classes may be the immense difference between arm US and LUS, which is more than between LUS classes. Although this method was described before with good results[18], it is questionable whether using arm US videos labeled as ID is a valid method for training the model. The reason ID was chosen to be one of the classes in the first place, was that it is desirable that the model is able to recognize images of poor quality. This could be caused by e.g. poor contact between the tissue and probe, by movement of the patient during the measurement or by imaging below the diaphragm. Although arm US videos contain some features similar to LUS images, such as bone and subcutaneous tissue, they are still very different from LUS videos and the validity may increase if the ID videos were more realistic to actual indeterminable videos during a LUS examination. This may also explain why the model performs high on ID videos, despite the relatively little amount of data. Based on the skewed data distribution, the expectation would be that the model would perform best on score 1 and second best at predicting score 0. However, it performs second best at predicting ID. It makes sense not to force the model to make a prediction in one of the four LUS scores and so to include ID, but as the model is only an assisting tool and will (for now) not be used as a diagnostic tool, excluding ID as a class should be considered.

Results from previous studies are divergent; Khan et al.[60] found the biggest classification errors in score 3 predicted as 2, 0 predicted as 2 and 2 predicted as 3. The agreement between their model and clinical experts was approximately 50%. Another study[61] showed that the accuracy for predicting score 3 was highest, followed by 1, 0 and 2. Roy et al.[12] demonstrated a weighted sensitivity of 49% $\pm$ 18, F1-score of 46 % $\pm$ 21 and precision of 55% $\pm$ 27. These are very similar to the average results of the current study. Considering their confusion matrices

they generally showed overestimation of the LUS scores. Dastider et al.[13] trained separate networks for a linear probe and a convex probe. The linear probe achieved the highest results with an average accuracy of 79%, sensitivity of 79%, specificity of 90% and F1-score of 79%. Regarding the accuracy and specificity it is very similar to the results of this study.
A major limitation in labeling the videos, was that we have noticed that the traditional LUS scoring system, varying from 0 to 3, may not be sufficient for our assessment. We have noticed that a broad range of aeration loss could be classified as a score 1; according to our definitions for each score in this study, no significant B-lines (0 to 2 lines) are defined as a score 0, whereas coalescent B-lines and/or large consolidations with less than 50% loss of aeration are defined as a score 2. That implies that everything in between is defined as a score 1, which turned out to be much. The wide variety in class score 1 is depicted in Figure 6.1. This can also be observed in the data distribution for the video-based classification; the amount of videos classified as a score 1 is disproportionate to the other scores. This may explain the high sensitivity for score 1, as statistically the network may predict most videos as score 1 and be correct. When considering Figure 5.2 relatively many videos having score 0 are predicted as score 1, while we noticed during labeling that most ambiguity was between score 1 and 2. A possible explanation may be variation in image characteristics, as a result of differences in gain and lack of gel use. Therefore, if we would score the images again, we could introduce score 1a and score 1b to take out these nuances in the score 1 classified images. A score 1b could then be assigned to images which do not apply for a score 2, but tend to. If these new scores are not to be introduced, it is recommended to collect more data of the other classes and reconsider the scoring method as in Table 4.1.



(a) (b) (c)

Figure 6.1: Examples of the wide variety in videos that have been labeled as score 1 by clinical experts. Frame (a) was labeled with score 1 because of subpleural consolidations without B-lines, (b) because of multiple B-lines and (c) because of the near coalescent B-lines (which were not coalescent enough for score 2).

Another possibility to improve classification results, would be to cluster classes based on technical similarities in the data. A suggestion is to group score 0 and score 1a images together,

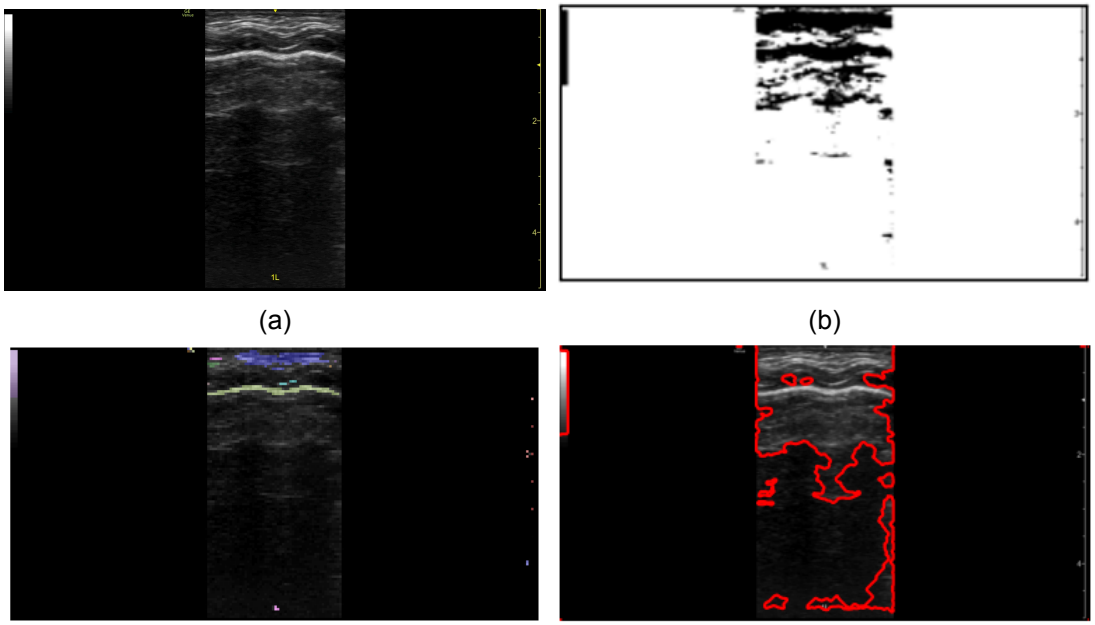
and score 1b and 2. Therefore the amount of classes would be reduced to 3 instead of 4. Not only could this improve the performance of the model, clinically it would also make more sense as the classes could then be correlated to clinical implications or actions more. A score 0 and 1a do not differ regarding clinical implications, while a score 1b and 2 could be an indication to start with or increase diuretics therapy. A score 3 requires consideration of the cause - consolidations or atelectasis - and the severity of the affected regions to effectuate treatment.
Another suggestion for future research is to investigate different classification methods as Roy et al.[12]. They first made frame by frame predictions, after which they either scored the entire video based on the maximal frame-level score, or the average frame-level score. They have described that the former strongly predicts towards higher scores, while the latter performs well on score 0 and poorly on all other scores. Applying similar methods require more elaborate network designs that are beyond the scope of this thesis.
Considering the results of the logistic regression model, we can conclude that a basic method does not suffice for the classification of LUS images, considering sensitivities and specificities of 0 and 100%. It is evident that the pixel count of manually segmented pathological features of LUS frames are not a good predictor for the LUS score. From the results in Figure 5.4 and Table 5.3 we can conclude that the model is able to predict when the frame does not have a score 0, 2 and 3, but has difficulties in predicting not having score 1. Again the opposite also counts, so the model can often predict a score 1 correctly. However, there are a lot of remarks to this method. First of all, the pixel count of segmentation masks was chosen as a feature to predict LUS scores by lack of an alternative, but we have not corrected for the extent of rib shadows or thickness of tissue above the pleurae. Therefore the absolute pixel count is by definition not a reliable feature. Also the data distribution was not equal and may have affected model training. Thereby it has demonstrated that the use of an RNN is more suitable for LUS score prediction than a simple logistic regression model.
As mentioned before, LUS is a skill that should be trained and therefore it is only understandable that there is a certain learning curve in both making images and scoring them. Therefore there is a possibility that there was a high intra-observer variability in the assessment of images. Additionally, we have not done any structural measurement or correction for the interobserver variability. All videos were labeled after consensus between at least two independent observers was reached, but we have no measurements of the independent scoring and variability. It is a possibility that the interobserver variability is already high, which makes it challenging for a deep learning network to find a generalizable pattern in the data. For example, Kumar et al.[10] demonstrated an interobserver reliability of 79% when determining the total B-line count per scan and 56% for lateral consolidations and 86% for posterior consolidations. The interobserver reliability was lowest for bilateral consolidations (28%).
Another point of discussion is the validity of data augmentation. Augmentation prevents the network from learning irrelevant features, such as the black space on each side of the image or the mark at the bottom indication the lung region. By flipping and rotating the image, these features are changed constantly and we prevent that the model tries to find a pattern in those features. Bright features such as A-lines and B-lines stay bright and therefore hopefully the models picks up those features. Previous work has shown that augmentation may improve model performance[42][43]. Although both flipping and rotation show different aspects of the dataset, the model is translationally invariant. Therefore it is recommended to not apply data augmentation to the dataset in future studies.
The results of the frame-based segmentation are high; average DSC's of 96% and 97% and MSE's of 0.03 of the two last saved models imply a high similarity between the desired output segmentation masks and the masks generated by the GAN. Another study[41] demonstrated an average DSC between 78% and 86% using transfer learning methods on pre-trained networks.

This is in accordance with other work with DSC's between 64% and 75%[12] and between 75% and 90%[62]. Compared to previous studies, we have found higher agreements between the output and desired segmentation masks. This could be explained by a large part of the segmentation masks being uninformative black space with several bright pixels. Even if the model fails to localize multiple B-lines, which clinically would have implications, the DSC would still be rather high. An alternative may be to not localize clinical features by segmentation, but by object detection with bounding boxes. In that case the amount of clinical features would be detected, rather than the amount of pixels per frame. As the LUS score is dependent on B-line counts, feature extraction is a common method in studies on DL for LUS assessment. However, B-lines vary in thickness and studies have shown that the underlying pathologies can be distinguished by B-line shapes[25]. For the B-line detection, semantic segmentation was applied. However, with this method each input frame only has one segmentation mask. Therefore the segmentation masks include all features: A-lines, B-lines and consolidations. Labeling with bounding boxes and training the model with that data may be a good method to distinguish the features. Although, B-line appearances are heterogeneous based on the etiology: B-lines generated by a fibrotic or inflammatory lung have a different appearance from those generated by cardiogenic edema and therefore B-line detection only is not sufficient and the appearance should be analyzed[25].

A disadvantage of frame-based segmentation, is that individual frames are a lot less informative than dynamic videos. When labeling the images, there was often doubt about clinical features, which would have been less in dynamic images. The images have been labeled conservatively, meaning that when in doubt features were not segmented. This was often the case for B-lines and small subpleural consolidations. To the contrary, A-lines were rather easy to segment, as they occur at specific distances from the previous lines. Therefore the model may have been trained very well on A-lines, but worse on B-lines. Despite the relatively easy labeling of A-lines, it seems that the model is able to predict the location and measurements of B-lines better. This can also be observed in Figure 5.6, where the model seems to generate segmentation masks with B-lines well, while it does not show any A-lines. As most studies have focused on the automatic segmentation of B-lines, there is no explanation described in literature. Whether or not to include A-lines in the segmentation masks should be considered. As the presence of many A-lines suggests well aerated lung parenchyma or pneumothorax, they provide valuable information. However, it would be more consistent to only include pathological features in the segmentation masks.

Also in labeling the segmentation masks we have no results for the interobserver variablity. A study[63] demonstrated the intraclass correlation coefficient (ICC) for clinical experts and a neural network when counting B-lines from LUS images. The agreement between their visual and automatic assessment was 79%, with an ICC among clinical experts between 62%-0.99%. For the automatic system the ICC was 49%-83%..

Initially we tried to label the images for the semantic labeling with different automatic segmentation methods, among which threshold-based, watershed and active contours segmentation, as can bee seen in Figure 6.2. However, it was difficult to apply the algorithms to all LUS images, as they were very heterogeneous and the clinical features were not evenly prominent in all images; some images had clear A- and/or B-lines, while in others the whole image was bright and features were difficult to extract. Moreover, the shadows of the ribs reduced the images to small regions with useful clinical information, which complicated segmentation. We found that the efforts and time necessary to automatically segment the images were not proportionate to the results, compared with manual labeling. Therefore we have made the choice to manually label all images.



(a) (b)

(c) (d)

Figure 6.2: Attempts to automatic segmentation of (a) a test frame, using (b) Otsu's threshold method, (c) watershed segmentation and (d) active contours.

Although the average results of the RCNN are very similar to the results of previous studies, the poor sensitivities in predicting score 0 and 2 make that the model is not ready to be used in practice yet. When the model will be used to predict scores, chances are high that the model will score the video as LUS score 1 or ID. It is recommended to first improve the model with aforementioned suggestions and then to introduce the model as a assisting rather than a decisive tool in clinical practice. This implies that the clinician should have moderate understanding of LUS interpretation and that the model can be used for confirmation or an additional control. The GAN that was developed for this thesis seems to be capable of segmenting important features from independent frames and with the aforementioned improvements, it is expected that the model can be used for both clinical implementation and education of unexperienced ultrasonographists. The model is able to localize important features from the frames and may provide valuable information for lung aeration assessment.
When both models have achieved performances that are sufficient for clinical implementation, LUS can more easily be used by multiple clinical professionals and hopefully the amount of CXR examinations can be reduced.

7 AKNOWLEDGEMENTS

There are many people who were part of my internship who I would like to thank, some people in particular. I would like to thank Can for brainstorming with me about the technical elements of this thesis and for helping me with the many programming frustrations. Anyone who is familiar with programming understands how incredibly valuable this is.
Many thanks to Sabien and Ariane who have educated me from the very beginning to be a confident medical professional. I have much enjoyed our rather long, but always fun consensus meetings. Special thanks to Sabien for not only her patience when the internship proposal was first rejected, but also for taking me under her wing, making sure I would find my place in the group and always being concerned about my well-being. I have gratefully used the office space (including coffee) she was more than willing to share for a year.
Thanks should also go to Nicole, who has witnessed all highs and lows of the past two years and made sure us students would not only get good grades, but actually would turn out to be great professionals. Thank you for gently pushing me out of my comfort zone every now and then.
I also had great pleasure of working with all colleagues who where patient with me and from whom I have learned everything I now know about intensive care. Naturally I am also more than grateful for my friends, family and everyone else who made sure I kept enjoying life outside of work during the past year.

REFERENCES

[1] M. Barile. Pulmonary edema: A pictorial review of imaging manifestations and current understanding of mechanisms of disease. *European Journal of Radiology Open*, 7, 2020.

[2] R. Claure-Del Granado and R.L. Mehta. Fluid overload in the icu: Evaluation and management. *BMC Nephrology*, 17, 2016.

[3] A.S. Messmer, C. Zingg, M. Müller, J.L. Gerber, J.C. Schefold, and C.A. Pfortmueller. Fluid overload and mortality in adult critical care patients-a systematic review and meta-analysis of observational studies. *Critical care medicine*, 48:1862–1870, 2020.

[4] D.S. Prough and C.H. Svensén. Perioperative fluid management. *Anesthesia and Analgesia*, pages 84–91, 2006.

[5] U.A. Tandircioglu, S. Yigit, B. Oguz, G. Kayki, H.T. Celik, and M. Yurdakok. Lung ultrasonography decreases radiation exposure in newborns with respiratory distress: a retrospective cohort study. *European Journal of Pediatrics*, 1:1–7, 2021.

[6] International Atomic Energy Agency. Radiation in everyday life. Available at `https://www.iaea.org/Publications/Factsheets/English/radlife`(05/10/2022).

[7] E.J. Hall and D.J. Brenner. Cancer risks from diagnostic radiology: The impact of new epidemiological data. *British Journal of Radiology*, 85, 2012.

[8] A. Ammirabile, D. Buonsenso, and A. Di Mauro. Lung ultrasound in pediatrics and neonatology: An update. *Healthcare (Switzerland)*, 9, 2021.

[9] E.J. Hall. Lessons we have learned from our children: cancer risks from diagnostic radiology. *Pediatric Radiology 2002 32:10*, 32:700–706, 2002.

[10] A. Kumar, Y. Weng, S. Graglia, S. Chung, Y. Duanmu, F. Lalani, K. Gandhi, V. Lobo, T. Jensen, J. Nahn, and J. Kugler. Interobserver agreement of lung ultrasound findings of covid□19. *Journal of Ultrasound in Medicine*, 40:2369–2376, 2021.

[11] J.J. Rouby, C. Arbelot, Y. Gao, M. Zhang, J. Lv, Y. An, W. Chunyao, D. Bin, C.S.V. Barbas, F.L.D. Neto, F.P. Caltabeloti, E. Lima, A. Cebey, S. Perbet, J.M. Constantin, H. Brisson, R. Deransy, C. Vezinet, P. Garçon, N. Kacem, D. Lemesle, A. Monsel, Q. Lu, O. Langeron, F. Gay, B. Lucena, L. Malbouisson, M.J.C. Carmona, J. Neves, P. Dalcin, G. Schettino, A. Biestro, D. Cristovao, and J. Salluh. Training for lung ultrasound score measurement in critically ill patients. *American Journal of Respiratory and Critical Care Medicine*, 198:398–401, 2018.

[12] S. Roy, W. Menapace, S. Oei, B. Luijten, E. Fini, C. Saltori, I. Huijben, N. Chennakeshava, F. Mento, A. Sentelli, E. Peschiera, R. Trevisan, G. Maschietto, E. Torri, R. Inchingolo,

A. Smargiassi, G. Soldati, P. Rota, A. Passerini, R.J.G. Van Sloun, E. Ricci, and L. Demi. Deep learning for classification and localization of covid-19 markers in point-of-care lung ultrasound. *IEEE Transactions on Medical Imaging*, 39:2676–2687, 2020.

[13] A.G. Dastider, F. Sadik, and S.A. Fattah. An integrated autoencoder-based hybrid cnn-lstm model for covid-19 severity prediction from lung ultrasound. *Computers in Biology and Medicine*, 132:104296, 2021.

[14] M. La Salvia, G. Secco, E. Torti, G. Florimbi, L. Guido, P. Lago, F. Salinaro, S. Perlini, and F. Leporati. Deep learning and lung ultrasound for covid-19 pneumonia detection and severity classification. *Computers in Biology and Medicine*, 136:104742, 2021.

[15] R.J.G. van Sloun and L. Demi. Localizing b-lines in lung ultrasonography by weakly supervised deep learning, in-vivo results. *IEEE Journal of Biomedical and Health Informatics*, 24:957–964, 2020.

[16] M.R. Panicker, Y.T. Chen, K.V. Narayan, and C. Kesavadas. An approach towards physics informed lung ultrasound image scoring neural network for diagnostic assistance in covid-19. *arXiv preprint arXiv:2106.06980*, 2021.

[17] A. Tripathi, A. Rakkunedeth, M.R. Panicker, J. Zhang, N. Boora, J. Knight, J. Jaremko, Y. Tung Chen, and K. Vishnu Narayan. Physics driven domain specific transporter framework with attention mechanism for ultrasound imaging. *arXiv preprint arXiv:2109.06346*, 2021.

[18] J. Born, N. Wiedemann, M. Cossio, C. Buhre, G. Brändle, K. Leidermann, J. Goulet, A. Aujayeb, M. Moor, B. Rieck, and K. Borgwardt. Accelerating detection of lung pathologies with explainable ultrasound image analysis. *Applied Sciences*, 11:672, 2021.

[19] S. Kulhare, X. Zheng, C. Mehanian, C. Gregory, M. Zhu, K. Gregory, H. Xie, J. McAndrew Jones, and B. Wilson. Ultrasound-based detection of lung abnormalities using single shot detection convolutional neural networks. volume 11042 LNCS, pages 65–73. Springer Verlag, 2018.

[20] A. van Oosterom and T. F. Oostendorp. *Echografie*. Reed Business, 2015.

[21] H.Y. Liang, X.W. Liang, Z.Y. Chen, X.H. Tan, H.H. Yang, J.Y. Liao, K. Cai, and J.S. Yu. Ultrasound in neonatal lung disease. *Quantitative Imaging in Medicine and Surgery*, 8:535, 2018.

[22] S. Bobillo-Perez, M. Girona-Alarcon, J. Rodriguez-Fanjul, I. Jordan, and M. Balaguer Gargallo. Lung ultrasound in children: What does it give us? *Paediatric Respiratory Reviews*, 36:136–141, 2020.

[23] L. Demi, T. Egan, and M. Muller. Lung ultrasound imaging, a technical review. *Applied Sciences*, 10:462, 2020.

[24] A. Wong, C. Kirkpatrick, A. Longmead, H. Venables, and P. Parker. Lung ultrasound, 2020.

[25] A.M. Musolino, P. Tomà, C. De Rose, E. Pitaro, . Boccuzzi, R. De Santis, R. Morello, M.C. Supino, A. Villani, P. Valentini, and D. Buonsenso. Ten years of pediatric lung ultrasound: A narrative review. *Frontiers in Physiology*, 12, 2022.

[26] G. Soldati, M. Demi, R. Inchingolo, A. Smargiassi, and L. Demi. On the physical basis of pulmonary sonographic interstitial syndrome. *Journal of Ultrasound in Medicine*, 35:2075–2086, 2016.

[27] G. Volpicelli, V. Caramello, L. Cardinale, A. Mussa, F. Bar, and M.F. Frascisco. Detection of sonographic b-lines in patients with normal lung or radiographic alveolar consolidation. *Medical science monitor : international medical journal of experimental and clinical research*, 14:122–128, 2008.

[28] F. Raimondi, N. Yousef, F. Migliaro, L. Capasso, and D. De Luca. Point-of-care lung ultrasound in neonatology: classification into descriptive and functional applications. *Pediatric Research 2018 90:3*, 90:524–531, 2018.

[29] B. Bouhemad, S. Mongodi, G. Via, and I. Rouquette. Ultrasound for "lung monitoring" of ventilated patients. *Anesthesiology*, 122:437–447, 2015.

[30] D.A. Lichtenstein. Lung ultrasound in the critically ill. *Annals of intensive care*, 4:1, 2014.

[31] N.J. Soni, R. Franco, M.I. Velez, D. Schnobrich, R. Dancel, M. I. Restrepo, and P.H. Mayo. Ultrasound in the diagnosis and management of pleural effusions. *Journal of Hospital Medicine*, 10:811–816, 2015.

[32] A. Saraogi. Lung ultrasound: Present and future. *Lung India : Official Organ of Indian Chest Society*, 32:250, 2015.

[33] M. Karthika, D. Wong, S.G. Nair, L.V. Pillai, and C.S. Mathew. Lung ultrasound: The emerging role of respiratory therapists. *Respiratory Care*, 64:217–229, 2019.

[34] MATLAB Simulink. What is deep learning? | how it works, techniques applications. Available at `http://https://www.mathworks.com/discovery/deep-learning.html` (02/09/2022).

[35] R. Sun. Optimization for deep learning: theory and algorithms. *arXiv preprint arXiv:1912.08957*, 2019.

[36] Y. Lecun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015.

[37] C. Igiri, I.C. Peace, A.O. Uzoma, and S. Abasiama Ita. Effect of learning rate on artificial neural network in machine learning erp system for oil and gas view project football result prediction system view project effect of learning rate on artificial neural network in machine learning. *Article in International Journal of Engineering Research*, 2021.

[38] C.C Aggarwal. *Neural Networks and Deep Learning*. Springer International Publishing AG, part of Springer Nature, 2018.

[39] M. Bilgehan and P. Turgut. Artificial neural network approach to predict compressive strength of concrete through ultrasonic pulse velocity. *Research in Nondestructive Evaluation*, 21:1–17, 2010.

[40] X. Zhang. *Machine Learning*. Springer, Singapore, 2020.

[41] D. Cheng and E.Y. Lam. Transfer learning u-net deep learning for lung ultrasound segmentation. *arXiv preprint arXiv:2110.02196*, 2021.

[42] C. Shorten and T.M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6:60, 2019.

[43] S. O'gara and K. Mcguinness. Comparing data augmentation strategies for deep image classification. 2019.

[44] Z. Alom, T.M. Taha, C. Yakopcic, S. Westberg, P. Sidike, S. Nasrin, M. Hasan, B.C. Van Essen, A.A.S. Awwal, and V.K. Asari. A state-of-the-art survey on deep learning theory and architectures. *Electronics 2019, Vol. 8, Page 292*, 8:292, 2019.

[45] J. Ker, L. Wang, J. Rao, and T. Lim. Deep learning applications in medical image analysis. *IEEE Access*, 6:9375–9379, 2017.

[46] E.W. Weisstein. Convolution. Available at `https://mathworld.wolfram.com/Convolution.html`(05/10/2022).

[47] S. Albawi, T.A. Mohammed, and S. Al-Zawi. Understanding of a convolutional neural network. volume 2018-Janua, pages 1–6. Institute of Electrical and Electronics Engineers Inc., 2018.

[48] P. Isola, J. Zhu, T. Zhou, and A.A. Efros. Image-to-image translation with conditional adversarial networks. 2016.

[49] Google Developers. Overview of gan structure. Available at `https://developers.google.com/machine-learning/gan/gan_structure` (17/08/2022).

[50] M. Loey, F. Smarandache, N. Eldeen, and M. Khalifa. Within the lack of chest covid-19 x-ray dataset: A novel detection model based on gan and deep transfer learning. 12:651, 2020.

[51] H. Ali and Z. Shah. Combating covid-19 using generative adversarial networks and artificial intelligence for medical images: Scoping review. *JMIR Medical Informatics*, 10:e37365, 2022.

[52] M.E. Karar, M.A. Shouman, and C. Chalopin. Adversarial neural network classifiers for covid-19 diagnosis in ultrasound images. *Computers, Materials and Continua*, 70:1683–1697, 2021.

[53] R. Dipietro and G.D. Hager. *Deep learning: RNNs and LSTM*. First edition, 2020.

[54] B. Kumaraswamy. *Neural networks for data classification*. Elsevier, 2021.

[55] B. Barros, P. Lacerda, C. Albuquerque, and A. Conci. Pulmonary covid-19: Learning spatiotemporal features combining cnn and lstm networks for lung ultrasound video classification. 2021.

[56] M. Tkachenko, M. Malyuk, A. Holmanyuk, and N. Liubimov. Label studio: Data labeling software. Available at `https://github.com/heartexlabs/label-studio`(15/07/2022).

[57] M. Yeung, E. Sala, C. Schönlieb, and L. Rundo. Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics*, 95:102026, 2022.

[58] C. Szegedy, W Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. pages 1–9. IEEE, 2015.

[59] X. Yi, E. Walia, and P. Babyn. Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58:101552, 2019.

[60] U. Khan, F. Mento, L. Nicolussi Giacomaz, R. Trevisan, A. Smargiassi, R. Inchingolo, T. Perrone, and L. Demi. Deep learning-based classification of reduced lung ultrasound data from covid-19 patients. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 69:1661–1669, 2022.

[61] W. Xing, C. He, J. Li, W. Qin, M. Yang, G. Li, Q. Li, G. Wei, W. Li, J. Chen, and D. Ta. Automated lung ultrasound scoring for evaluation of coronavirus disease 2019 pneumonia using two-stage cascaded deep learning model. *Biomedical Signal Processing and Control*, 75:103561, 2022.

[62] B. Frey, L. Zhao, T.C. Fong, L. Bell, M.A.L. Bell, and M.A. Lediju Bell. Multi-stage investigation of deep neural networks for covid-19 b-line feature detection in simulated and in vivo ultrasound images. volume 12033, pages 52–59, 4 2022.

[63] J. Short, C. Acebes, G. Rodriguez-De-Lema, G.M.C. La Paglia, M. Pavón, O. Sánchez-Pernaute, J.C. Vazquez, F. Romero-Bueno, J. Garrido, and E. Naredo. Visual versus automatic ultrasound scoring of lung b-lines: reliability and consistency between systems. *Medical ultrasonography*, 21:45–49, 2019.