

UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE CIENCIAS FÍSICAS

DEPARTAMENTO DE ESTRUCTURA DE LA MATERIA, FÍSICA
TÉRMICA Y ELECTRÓNICA



TRABAJO DE FIN DE GRADO

Código de TFG: ETE20

Imagen por Ultrasonidos

Ultrasound Imaging

Supervisor/es: Joaquín López Herraiz

Ignacio Ortega Alonso

Grado en Física

Curso académico 2020-2021

Convocatoria de febrero

B-line scoring and detection in lung ultrasound

Resumen:

La ecografía o ultrasonografía es una técnica de imagen ampliamente extendida con múltiples aplicaciones en medicina e industria. Se fundamenta en la emisión de ondas de sonido de alta frecuencia y la posterior recepción de sus ecos. Midiendo tiempos de retraso y cambios en la amplitud de presión de la onda recibida, se puede inferir la estructura del objeto escaneado y sus propiedades acústicas. A pesar de su naturaleza poco apropiada para la ecografía, los pulmones también se pueden examinar mediante esta técnica. En particular, ciertos artefactos de la imagen conocidos como 'líneas B' son indicativos de neumonía y otras patologías. Mediante un algoritmo de aprendizaje automático entrenado sobre una colección de exploraciones de pulmón evaluadas, se pueden detectar líneas B en función de una escala de puntuación. Los resultados se podrían mejorar con cambios en el sistema de puntuación y en el enfoque elegido.

Abstract:

Ultrasound sonography or ultrasound is a widely used imaging technique with several medical and industrial applications. It is based on the emission of high-frequency sound waves and the reception of their returning echoes. Measuring time delays and changes in pressure amplitude of the received waves, the scanned object's structure and acoustical properties can be inferred. Despite acoustical properties unsuited for ultrasound imaging, lungs can also be examined. Image artifacts known as B-lines are indicative of pneumonia and similar abnormalities. Using a collection of assessed lung scans, a machine learning algorithm is designed to detect B-lines based on a scoring scale. The results could be improved with changes in the scoring system and the approach taken.

Contents

1	Introduction	3
1.1	Physical principles of ultrasound imaging	3
1.2	Transducers, beam properties	4
1.3	Image formation	6
1.4	Lung Ultrasound and artifacts	6
2	Objectives	7
3	Methods	8
3.1	Dataset	8
3.2	Video segmentation	8
3.3	Polar format	9
3.4	B-line scoring	10
3.5	Logistic regression	11
4	Results	12
5	Discussion	14

1 Introduction

1.1 Physical principles of ultrasound imaging

Medical ultrasound allows to obtain images of different body tissues thanks to the controlled emission of high-frequency sound waves and the reception of their reflections. Ultrasound probes are equipped with piezoelectric transducers that vibrate when excited with high-voltage electric pulses. This vibration is then transmitted through the body as a pressure wave whose behaviour is influenced by the mechanical properties of the tissues. A fraction of the energy emitted by the probe is reflected on tissue interfaces and received by the transducers. The transducers then vibrate and produce an electric signal that is processed in order to create an image of the examined region.

Sound waves can be understood as periodic variations of pressure on a propagation medium. The rate at which the pressure varies (frequency f) and the spatial distance between neighbouring points with equal pressure (wavelength λ) are related through the speed of sound $c = \lambda f$. The speed of sound c depends on the elastic properties of the propagation medium, and it can be expressed as a function of bulk modulus B and density ρ by:

$$c = \sqrt{\frac{B}{\rho}}$$

In soft tissue, sound travels at an average speed of 1540ms^{-1} , but it can be as slow as 600ms^{-1} in the lungs, or as fast as 3200ms^{-1} in bones (see table 1).

Another property of a medium which is related to the speed of sound is acoustic impedance, defined as:

$$Z = \rho c$$

Differences in acoustic impedances of adjacent tissues are responsible for the reflection of sound waves. More exactly, the ratio between the pressure amplitudes of the reflected (P_r) and incident (P_i) wave can be expressed as

$$\frac{P_r}{P_i} = \frac{Z_2 - Z_1}{Z_2 + Z_1}$$

where Z_1, Z_2 are the acoustic impedances of the proximal and distal tissues respectively. An intensity reflection coefficient can also be defined as:

$$R_I = \frac{I_r}{I_i} = \left(\frac{P_r}{P_i}\right)^2$$

making use of the fact that intensity is proportional to the squared power of pressure amplitude. For some interfaces like fat-muscle this coefficient is very small, meaning that most of the energy carried by the wave keeps propagating. For others, such as muscle-air almost all the energy is reflected.

When the size of the objects responsible for reflection is similar or smaller than the wavelength, sound is reflected in all directions or scattered. The amount of scattered energy

	$\rho(kg \cdot m^{-3})$	$c(m \cdot s^{-1})$	$Z(kg \cdot m^{-2} \cdot s^{-1})$
Air	1.2	330	400
Lung	300	600	$1.8 \cdot 10^5$
Soft tissue	1050	1540	$1.6 \cdot 10^6$
Bone	1650	3200	$5.3 \cdot 10^6$

Table 1: Typical values of density, speed of sound and acoustical impedance [1]

depends on acoustic impedance differences, but also on the texture of the tissue and the wavelength.

Scattering, together with tissue absorption, are responsible for the attenuation or loss of energy of the propagating waves. Different tissues have different attenuation coefficients. They are typically measured in units of dB/cm and are proportional to frequency. In other words, waves with longer wavelength can penetrate deeper into the body. [4]

1.2 Transducers, beam properties

Ultrasound waves are produced by transducers made of piezoelectric materials. Because of the way electric dipoles are arranged in this kind of materials, they vibrate when a voltage difference is applied to their surfaces. Similarly, when an external force makes them contract or expand, a difference in voltage appears between their surfaces. This property allows them to operate both as transmitters and receivers of sound waves. An important property of a transducer is its Q factor, which is defined as:

$$Q = \frac{f_0}{\Delta f}$$

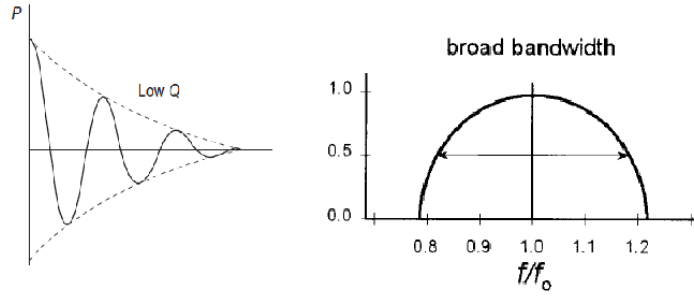
where f_0 is the resonant frequency of the transducer and Δf is its bandwidth, that is, the range of frequencies around f_0 where energy is also emitted. These frequencies are introduced by dampening the vibration of the piezoelectric and thus making the emitted pulse shorter. Transducers that are lightly damped show a narrow bandwidth (high Q factor) and a long spatial pulse length, whereas heavily damped transducers have broad bandwidths (low Q factor) and short spatial pulse length (figure 1).

Several piezoelectric elements are arranged in transducer arrays and operate together to produce an ultrasound beam. The shape of the beam is the result of interference between every individual emitter. It is characterized by two regions: a near field where the beam converges and a far field where it diverges. The length of the near field (or in other words, the focal distance) is related to the effective diameter of the transducer and its frequency by:

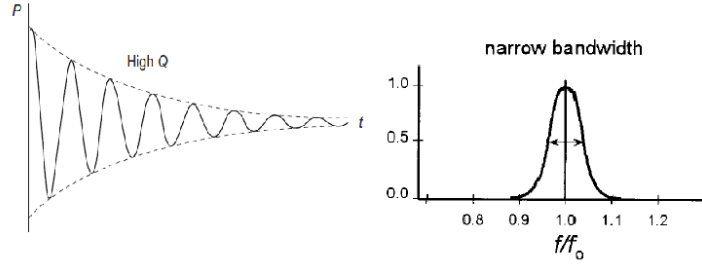
$$l \propto \frac{d^2}{\lambda}$$

That is, wider transducers and higher frequencies produce longer focal distance.

The ability to resolve objects in the acquired image is determined by a series of parameters of the beam and transducers. In the axial direction (direction of the beam), resolution is limited by the spatial pulse length (SPL). Two returning echoes can only be resolved if the distance between them is greater than the SPL. Therefore, for two reflectors to be distinguished they cannot be separated by a distance greater than half the SPL (figure 2).



(a) Low Q, broad bandwidth, short pulse



(b) High Q, narrow bandwidth, long pulse

Figure 1: Low Q and high Q: spectrum and pulse comparison [4] [1]

Shorter SPL can be achieved with shorter wavelength or fewer cycles per pulse, i.e. lower Q factor. Typical axial resolution values are of $0.5mm$. Because SPL is constant along the beam, axial resolution is independent of depth. In the perpendicular direction resolution is determined by the diameter of the beam. For this reason lateral resolution is best at the focal zone and it decreases in the near and far fields. Some transducer arrays can produce beams of variable focal distance, creating an extended focal zone where lateral resolution is optimal.

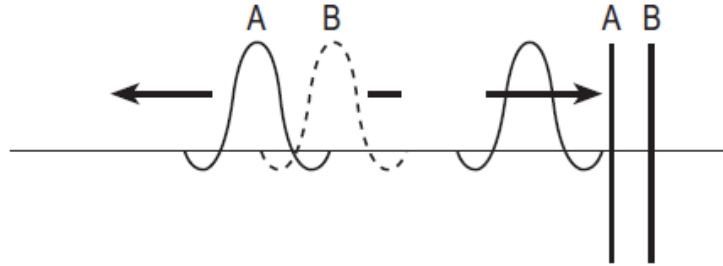


Figure 2: Axial resolution: separation between reflectors A and B must be must be smaller than $\frac{1}{2}SPL$ [1]

1.3 Image formation

In order to generate a two-dimensional ultrasound image several steps have to be followed. First transducers are set to 'transmit mode' to produce sound pulses using a high voltage spike. After the pulse is emitted, they are set to 'receive mode' in order to capture vibrations of returning echoes. Echoes received earlier correspond to reflections taking place close to the transducer, while later echoes are produced by deeper reflectors. The returning echoes are very weak and can only induce small voltages in the transducers that need to be amplified. In addition, later echoes are significantly weaker than earlier ones due to attenuation experienced by the sound wave throughout its round-trip. To compensate for this an additional time-dependent amplification known as time gain compensation (TGC) is introduced. The outcome of this process is a signal containing information of the scanned tissues along the beam direction, where amplitude is proportional to acoustic impedance differences and time corresponds to depth. Assuming an average speed of sound of $1540m/s$ along the beam path, time and depth are related through

$$D = \frac{ct}{2}$$

where the 2 factor accounts for the round-trip distance. After a certain time interval, transducers are set to emit the next pulse and the whole process is repeated several times per second with different focus parameters to achieve a two-dimensional image.

1.4 Lung Ultrasound and artifacts

As with any other imaging device, ultrasound cannot reproduce anatomy with unlimited accuracy. Artifacts are deviations from the actual aspect of the scanned objects due to the intrinsic working principles of the device. Lungs are particularly prone to exhibit artifacts when probed with ultrasound, but valuable information can still be obtained from its examination when the origin of these artifacts is well understood. A typical lung ultrasound image shows the following features. First, a very bright and approximately horizontal line can be seen in the upper half of the picture, corresponding to the pleura. The pleura is a thin membrane inside the chest cavity where lungs are contained. When probing perpendicular to the ribs, these can be also seen as dark regions above the pleura projecting shadows downwards. Due to the high difference of acoustic impedance between the soft tissue above the pleura and the air-filled lung below it, almost all the energy carried by sound waves is reflected, making observation of deeper tissues impossible. A common artifact that can be observed in the region below the pleural line is A-lines (figure 3.b). A-lines appear like equally distanced bright lines parallel to the pleural line and they can be explained as a reverberation artifact. Due to the high reflectivity of the pleura, echoes can bounce back and forth several times between it and the transducer surface (figure 3.a). Because of their delay with respect to the emitted pulse, these echoes are interpreted as multiple objects located further away from the transducer [2].

Another reverberation artifact that can be commonly observed in lung ultrasound is B-lines (figure 3.c). Whereas A-lines are usually present in healthy lungs, B-lines are correlated with interstitial syndrome, a kind of disease where inflammation or accumulation of fluid makes thin tissues within the lung called interstitium to thicken. This can be observed

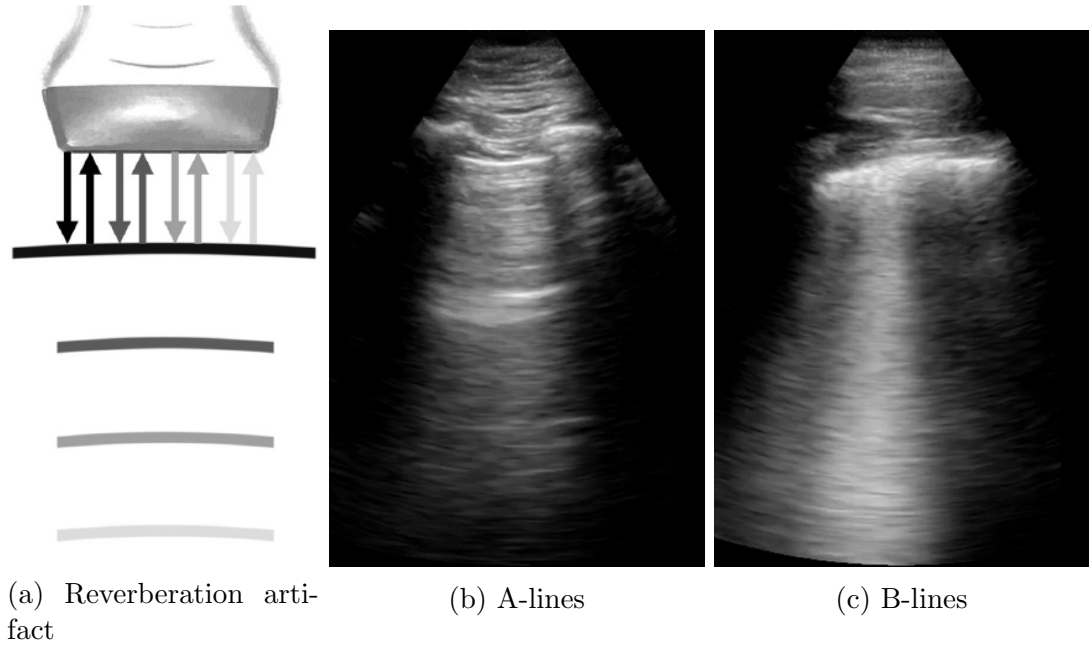


Figure 3: Common ultrasound artifacts found in lungs

for example in pneumonia patients. While the exact origin of this artifact is not clearly understood, it is thought that it is originated from reverberation between closely located high reflective surfaces, like air-liquid. Because of the proximity between reflectors the resulting echoes are not distinct, but rather appear as a bright and diffuse vertical line stretching from the pleural line downwards. During the current pandemic, lung ultrasound has proved to be a reliable method for diagnostic and monitoring of Covid-19 pneumonia through the assessment of B-lines and similar phenomena.

2 Objectives

The first goal in the present work has been to understand the physical principles of ultrasound imaging and to get familiarized with the technology involved. The focus of the study has been in lung ultrasound, its findings and their physical origin. In addition to this, current applications of artificial intelligence to medical physics have been explored, with an emphasis in lung ultrasound.

The ULTRACOV project (ultracov.daselsistemas.com) aims to design an ultrasound imaging device with artificial intelligence tools for the detection and monitoring of Covid-19 pneumonia. We attempt here to design a tool capable of processing a collection of videos, assess them on a particular scoring system and automatically detect B-lines. All steps of the process have been coded using Python (<https://github.com/igortega/B-lines>).

3 Methods

3.1 Dataset

In order to train our model we have made use of a collection of 280 ultrasound videos of lungs. The videos correspond to a single patient throughout a 40 day period of regular examinations monitoring the evolution of Covid-19 pneumonia. They were taken and labelled according to B-lines presence by an experienced physician (Dr. Yale Tung Chen - Hospital Universitario Puerta del Hierro) as part of the ULTRACOV project. All videos were taken with a Butterfly iQ probe with settings of scan depth ranging from 7 cm to 14 cm. Most of the scans belong to regions of the chest where lung is the only visible organ, but on a total of 78 videos other organs such as liver or spleen show up. An image resolution of 1080×632 pixels and framerate of 21 fps is constant across the whole database. Video durations vary widely, ranging from 1 second to 20 seconds of length, with an average of 6 seconds per video.

Each video was labelled depending on whether B-lines were present or not at any instant throughout the video. In other words, videos labelled as 'no B-lines' do not exhibit them on any of their frames whereas videos labelled as 'B-lines' do, at least on a fraction of their frames but not necessarily in all of them. This is due to the natural motion of the lungs while breathing, as well as motion of the probe and changes in direction of observation. Because of this, it is necessary to separate videos in smaller sections and isolate those corresponding to B-lines presence. A total number of 93 videos out of 280 are labelled as 'B-lines', which adds up to a 33% frequency of occurrence.

The frames making up each video were extracted on separate directories and cropped to remove unnecessary details for the later analysis. As a result, a collection of gray-scaled images sized 905×632 pixels and 8 bits of depth (i.e., brightness values ranging from 0 to 255) was obtained.

3.2 Video segmentation

Following the extraction and cropping of frames, every image set corresponding to a video has to be divided into smaller sets in order to simplify the analysis and isolate potential B-lines. To achieve this, a K-means clustering algorithm implemented in the Scikit-learn toolkit was used [7]. Clustering algorithms are a kind of unsupervised learning algorithms whose aim is to subdivide a dataset into separate groups or clusters in such a way that data points belonging to the same cluster are as close as possible to each other. K-means is one of the simplest of these algorithms. When applied to a two-dimensional dataset the result looks like the following (figure 4) [6]. Each cluster is represented by a cluster center or 'centroid', which corresponds to the average of the data points belonging to that cluster. Each data point is assigned to the cluster whose centroid is the closest. Different configurations are computed one after another until the total sum of distances within clusters is minimized, also known as 'inertia' parameter:

$$I = \sum_j^N \sum_{x_i \in C_j} \|\vec{x}_i - \vec{\mu}_j\|^2$$

where N is the total number of clusters, $\vec{\mu}_j$ is the 'centroid' of the j -th cluster C_j and \vec{x}_i are data points.

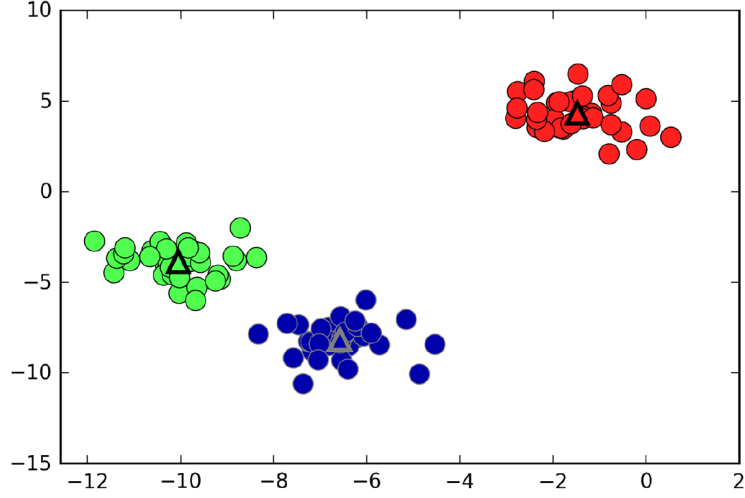


Figure 4: K-Means: each cluster (red, green, blue) is represented by its centroid (triangles) [6]

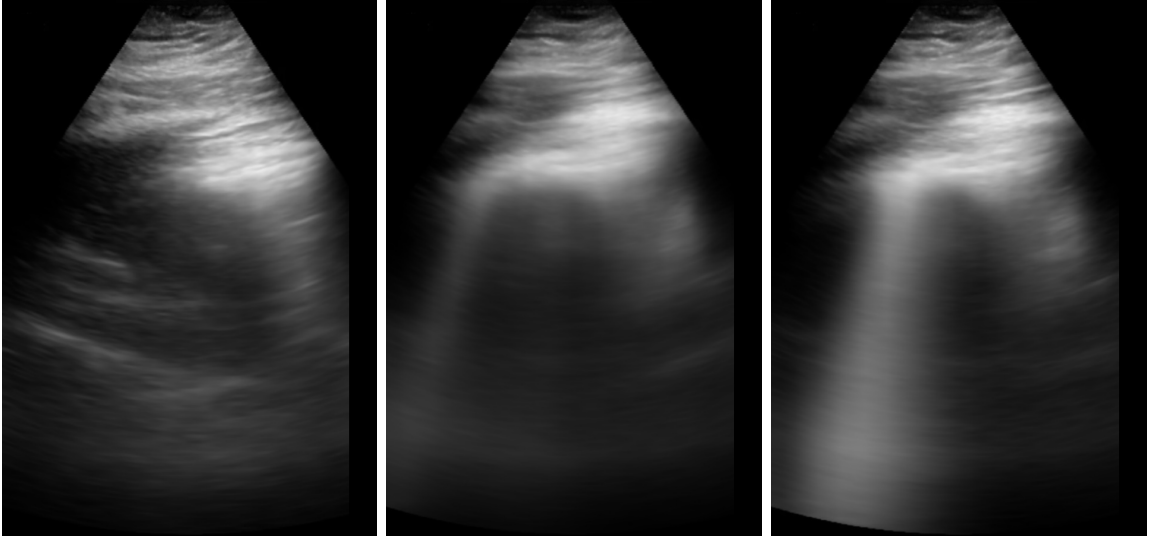


Figure 5: A set of three centroids representing clusters of frames of a single video

When working with images, instead of two dimensions each data point \vec{x}_i belongs to an n -dimensional space with n being equal to the number of pixels in the image ($905 \times 632 = 571960$ in our case). The resulting cluster centroids can be reshaped to the original dimensions to obtain averaged images of each cluster of frames. For every video in our collection three clusters have been computed with the purpose of obtaining a simplified representation of the video and isolating fragments of it that may exhibit B-lines (figure 5).

3.3 Polar format

Computing 3 cluster centroids per video greatly simplified the original dataset to a collection of images where each video is represented by three images. Before performing any analysis on the samples we have to make our dataset as homogeneous as possible. Once every video has been narrowed down to three images regardless of its duration, there are still

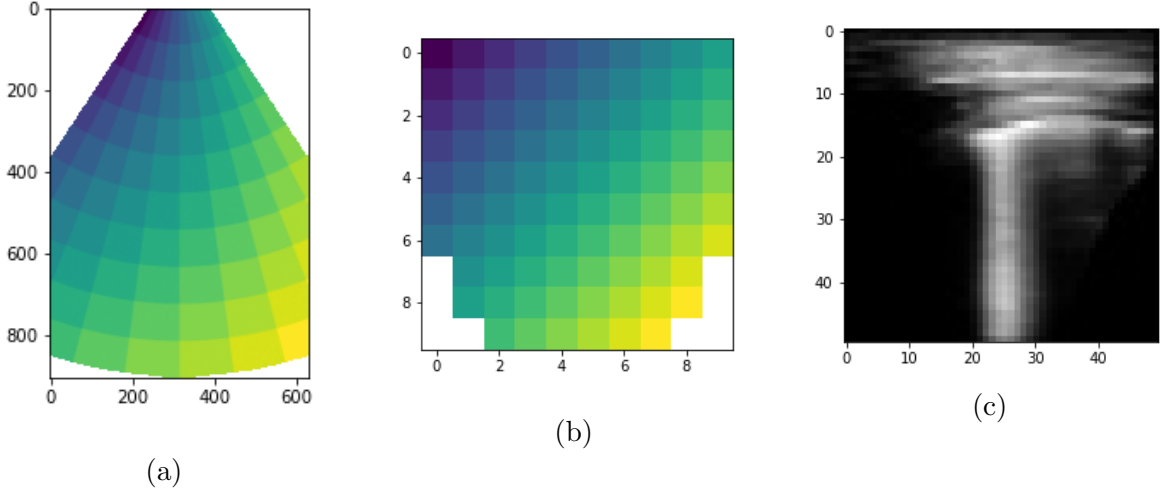


Figure 6: Reshaping of images into polar coordinates

differences across samples with respect to their maximum scan depth. These differences manifest themselves as changing aperture angles of the circular sector shapes of the images. Images with a shallower scan range show a wider aperture angle while those with a deeper range show narrower angles. In order to make these differences less apparent across the dataset we perform a reshaping of the images to a 'polar coordinates' format. To achieve this, we first manually measured the aperture angle on the image for each scan depth value (7, 8, 9, 10, 13 and 14 cm). Next, we divided the circular sector into several regions of equal radial and angular coordinates (figure 6.a). Finally, we took the mean value of brightness on each region and assigned it to the reshaped image (figure 6.b). By doing so, we obtained a rectangular-shaped image of N -rows by M -columns. In particular, we have chosen $N = M = 50$ (figure 6.c).

This reshaping of the images also facilitates further calculations, as we have greatly reduced the size of the images. It also helps identify B-lines more easily, as they should now appear as vertical lines whereas originally they could take diagonal shapes.

3.4 B-line scoring

We now describe a scoring system used to evaluate the likeliness of an image to show B-lines. Such a scoring system must take into account known features of B-line pattern: bright, uniform vertical shapes extending from the pleural line to the lower margin of the image. Images exhibiting B-lines should generate high scores on this scale, whereas those who show none should produce lower scores. The final score of a video is taken as the maximum score of all of its three centroids. All scorings are performed on the images in polar format. In this work three scoring scales have been used:

- **Total mean brightness:** average intensity is calculated for every column on the image. The highest value obtained is chosen as the outcome. Values are normalized to the peak brightness of the image.

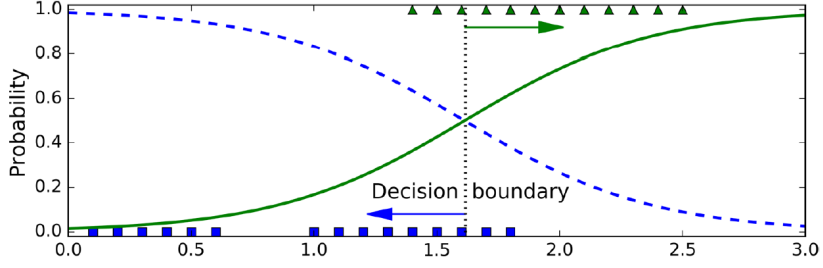


Figure 7: Example of a logistic function fit to a set of data. Values above the decision boundary are assigned to class 1. [5]

- **Lower quarter mean brightness:** average intensity is calculated for every column on the quarter section that is further away from the transducer. This should highlight the fact that B-lines extend down to the lowest margin of the image, while other artifacts fade away faster. Values are normalized to the peak brightness of the image.
- **Fraction of column above half-maximum:** maximum brightness is calculated across every column. Next, the number of pixels with intensity above half of this value is counted on every column, and its ratio to the total length of the column is taken as the outcome.

Once scored, every video should yield a set of three values corresponding to each scoring scale. All values are in the range $[0, 1]$ due to the normalization chosen. It is important to note that other scoring systems could be used. The one we present has been chosen because of its simplicity and easy interpretation.

3.5 Logistic regression

If our scoring system correlates well with the presence of B-lines, we should be able to train a machine learning algorithm that is able to predict the presence or absence of B-lines in a video based on its score. We have chosen a logistic regression implementation in Scikit-learn toolkit [7]. Logistic regression is a kind of supervised learning algorithm that assigns a data point to a class where it is most probable to belong to. It takes advantage of logistic functions, also known as sigmoids, by taking a continuous variable as an input and yielding a probability value between 0 and 1 as output.

$$f(x) = \frac{1}{1 + e^{-x}}$$

Probabilities below or above 0.5 correspond to one or the other class. The value of the input variable corresponding to a probability of 0.5 is known as the decision boundary (figure 7). In this case, three values per sample are fed to the algorithm as input, corresponding to each scoring scale, and two different values make up the output, corresponding to the probabilities of presence or absence of B-lines.

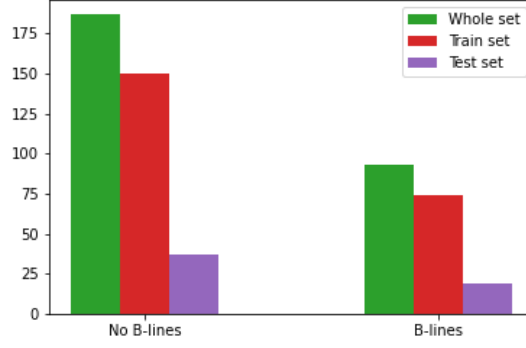


Figure 8: B-lines/no B-lines ratio of occurrence must be constant across all sets

4 Results

Before training the logistic regression algorithm, our dataset has been split into two different sets. The first one is the training set, i.e. the collection of labelled videos that will be fed to the algorithm. The second one is the test set, on which the algorithm’s prediction will be evaluated. Ratios of 0.8 and 0.2 out of the total dataset have been chosen respectively for the training and test sets. In other words, the algorithm is trained on 224 out of the total 280 samples, and it is tested on the remaining 56 samples. Both subsets must contain a similar ratio of samples labelled as B-lines as in the original dataset in order to remain representative (figure 8).

Scores of all videos in the training set have been calculated and represented against each other in figure 9. All of the three scoring scales are approximately correlated, as expected. That means that those videos who score high in one scoring scale also show high scores in the other scales, i.e. the scoring system is coherent. However, there is much overlap among scores of videos which show B-lines and those which do not. In other words, in this representation there is not a clear boundary between each class of data points.

A comparison between the algorithm’s prediction and the true values is summarized in table 1 in what is known as a ‘confusion matrix’. Elements in the diagonal of the matrix are correct predictions (true positives or true negatives), while elements out of the diagonal are wrong predictions (false positives or false negatives). The accuracy of the prediction is defined as the total ratio of correct predictions:

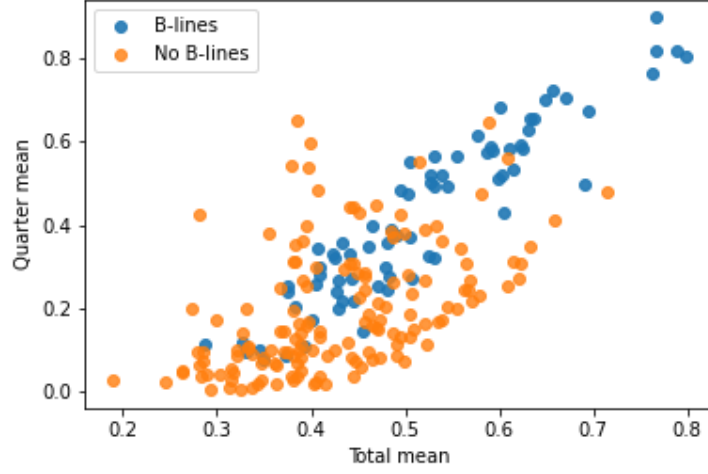
$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

With 39 correct predictions out of 56 samples the accuracy of the algorithm is $ACC = 70\%$.

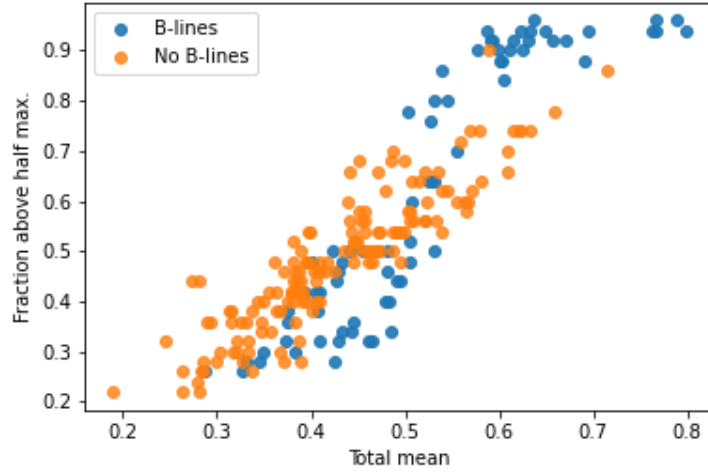
The ratio of correctly identified positives is measured by sensitivity or ‘true positive rate’ (TPR) and it is defined as:

$$TPR = \frac{TP}{TP + FN}$$

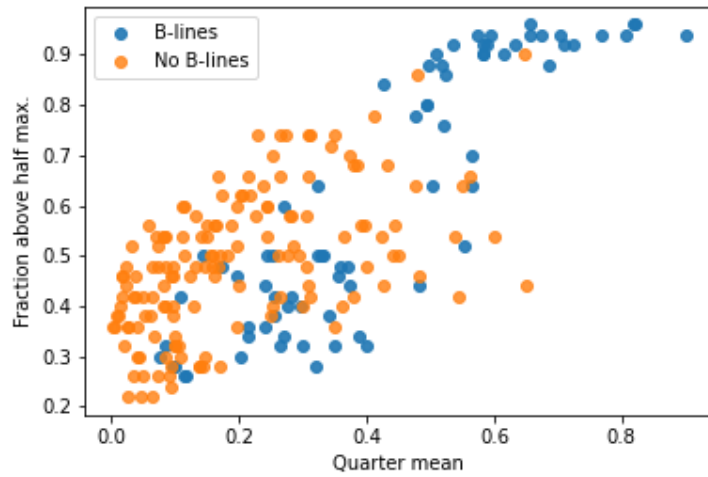
With only 5 predicted B-lines out of the existing 19, the achieved sensitivity is $TPR = 26\%$.



(a) Lower quarter mean brightness against total mean brightness



(b) Fraction above half-max against total mean brightness



(c) Fraction above half-max against lower quarter mean brightness

Figure 9: Score distribution of the training dataset in all combinations of the three scoring scales. True labels are shown (B-lines / no B-lines)

		Truth	
		B-lines	No B-lines
Prediction	B-lines	5	3
	No B-lines	14	34

Table 2: Confusion matrix: (left to right, top to bottom) true positives (TP), false positives (FP), false negatives (FN), true negatives (TN).

Conversely, the amount of true negatives detected is measured by specificity or 'true negative rate' (TNR):

$$TNR = \frac{TN}{TN + FP}$$

which in our case sums $TNR = 92\%$, with 34 out of 37 detected.

Finally, precision or 'positive predictive value' (PPV) is defined as the ratio of true positives to all predicted positives:

$$PPV = \frac{TP}{TP + FP}$$

Of a total of 8 predictions of B-lines, 5 are true positives, which implies a precision of $PPV = 63\%$.

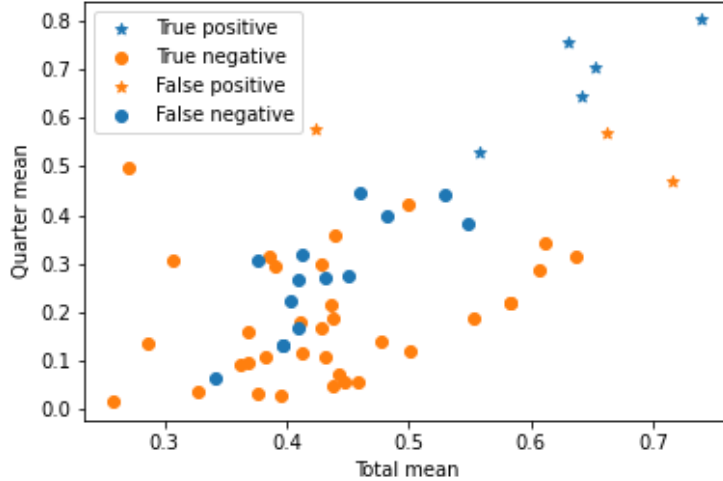


Figure 10: Score distribution and predictions on test set

5 Discussion

The proposed algorithm has shown a specificity of 92% in the identification of B-lines, meaning that very few false positives were produced. However, a sensitivity of only 26% has been achieved, implying a great number of false negatives. This can be explained by the score

distribution of the dataset (figure 9). It can be seen that there is a great overlap between data points belonging to different classes. Therefore, the algorithm cannot determine a clear boundary between them. This also explains the low accuracy obtained (70%) compared to the original distribution of B-lines (33%/66%). An algorithm that predicts every sample not to show B-lines would show an accuracy of 66%.

There are different explanations for the overlap in the score distribution. On one hand, the source of error can be in the scoring system itself. For example, the 'quarter mean' scale may not always be indicative of B-lines. When organs such as the spleen appear on a video, the brightness on the lower half of the image increases (see figure 11.a) even if there are no B-lines. This could cause a false positive. 'Total mean brightness' and 'fraction above half maximum' may also induce false positives due to the high intensity region above the pleural line, specially on the short range images (figure 11.b). Therefore, changes on the scoring system could improve the accuracy of the algorithm.

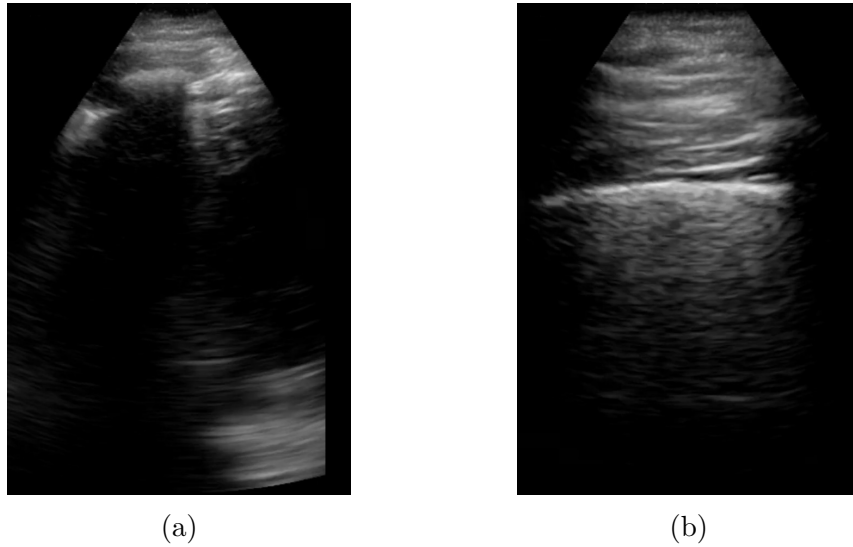


Figure 11: Examples of potential false positives with the proposed approach (a) Presence of organs such as the spleen may be mistaken for B-lines. (b) Short-range images with no B-lines may yield a high total mean brightness.

On the other hand, our representation of videos may not be accurate enough. Although in some cases at least one of the centroids does manage to isolate occurring B-lines (see figure 5), in others they do not, inducing false negatives. It could be because the B-lines appear very briefly on just a few frames or because of the swinging motion caused by breathing. This could be solved by increasing the number of centroids per video, or by manually selecting the target frames.

A similar approach has been taken with better results by Brattain et al. [3]. In their study, pleural lines were removed from the images before evaluating the presence of B-lines, which might facilitate uniformity across the dataset. The duration of the videos was also shorter, with a constant 60 frames per video. Another approach by van Sloun [8] used deep learning techniques, achieving accuracy of up to 90%.

References

- [1] Penelope J Allisy-Roberts and Jerry Williams. *Farr's physics for medical imaging*. Elsevier Health Sciences, 2007.
- [2] Michael Baad, Zheng Feng Lu, Ingrid Reiser, and David Paushter. Clinical significance of us artifacts. *Radiographics*, 2017.
- [3] Laura J Brattain, Brian A Telfer, Andrew S Liteplo, and Vicki E Noble. Automated b-line scoring on thoracic sonography. *Journal of Ultrasound in Medicine*, 2013.
- [4] Jerrold T Bushberg and John M Boone. *The essential physics of medical imaging*. Lippincott Williams & Wilkins, 2011.
- [5] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2019.
- [6] Andreas C Müller and Sarah Guido. *Introduction to machine learning with Python: a guide for data scientists*. " O'Reilly Media, Inc.", 2016.
- [7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.
- [8] Ruud JG van Sloun and Libertario Demi. Localizing b-lines in lung ultrasonography by weakly supervised deep learning, in-vivo results. *IEEE journal of biomedical and health informatics*, 2019.