Interpretable Prediction of Pulmonary Hypertension in Newborns using Echocardiography

Master Thesis

Author(s):
Ragnarsdottir, Hanna

Publication date:
2022

Permanent link:
https://doi.org/10.3929/ethz-b-000539940

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Interpretable Prediction of Pulmonary Hypertension in Newborns using Echocardiography

Master's Thesis

Hanna Ragnarsdóttir

March 21, 2022

Advisors: Prof. Dr. Julia Vogt, Dr. Ece Oezkanelsen, MSc. Laura Manduchi
Department of Computer Science, ETH Zürich

Abstract

Pulmonary hypertension (PH) in newborns is a rare but complex condition, associated with multiple diseases contributing to morbidity and mortality. Accurate and early detection of PH and the classification of its severity is crucial for appropriate and successful treatment. However, human assessment of PH using echocardiography, the primary diagnostic tool in pediatrics, is both expertise-demanding and time-consuming. Furthermore, little effort has been directed towards automatic assessment of PH using echocardiography, and the few proposed methods only focus on binary PH classification on the adult population. In this work, we propose a robust and interpretable deep learning approach to predict and classify the severity of PH in newborns, by utilising spatio-temporal patterns of the ultrasound videos from multiple views of the heart. To the best of our knowledge, this is the first work on multi-view video-based automated assessment of PH in newborns. Our results show a mean F1-score of 0.84 for severity prediction and 0.92 for binary detection using 10-fold cross-validation. To increase the clinical usability of our method, we complement our predictions with saliency maps that highlight how the learned model focuses on clinically relevant cardiac structures. We show that these learned localization maps align with how clinicians subjectively assess PH.

Contents

Chapter 1

Introduction

Pulmonary hypertension (PH) in newborns is a rare but complex condition, associated with multiple diseases contributing to morbidity and mortality. Accurate and early detection of PH and the classification of its severity is crucial for appropriate and successful treatment. However, human assessment of PH using echocardiography, the primary diagnostic tool in pediatrics, is both expertise-demanding and time-consuming, making early assessment difficult [24, 65]. Thus, there is a clear need for an automatic and streamlined method to assist clinicians in the assessment of PH in newborns. With little effort being directed towards automatic approaches for PH diagnostics, this need is not being met. The few existing methods for PH prediction are only proposed for the adult population and do not assess the PH severity nor propose methods to explain the predictions [37, 75].

The goal of this thesis is to meet the need for an explainable and automatic PH diagnostics tool, suitable for the use of pediatric cardiologists. Specifically, we aim to identify and create a robust deep learning method for automatic PH estimation of newborns, using heart ultrasound videos (echocardiography). Echocardiography is one of the most common and growing PH diagnostic tools due to its low-cost and non-invasive technology, which makes it an ideal choice for pediatrics [17, 47]. We are interested in not only predicting the existence of PH, but also in classifying its severity, as the appropriate PH treatment is determined based on the severity [12, 19]. We furthermore strive to explore the importance of the various factors contributing to the solution, including the effects of known deep learning techniques, such as data augmentation and regularisation, as well as the effects of various domain-specific factors. These include the addition of more than one ultrasound video per patient, from different views of the heart, as well as the inclusion or exclusion of the temporal domain of the videos to the solution. Finally, we seek to understand and explain the predictions with regard to cardiac structures and features, both to increase our own understanding of the method, and to increase the trust in the automatic prediction, ensuring clinical usability.

We propose four different approaches for PH assessment; a single-view spatial approach, a single-view spatio-temporal approach, a multi-view spatial approach and a multi-view spatio-temporal approach. We show with empirical assessments that the spatio-temporal multi-view approach has the greatest predictive power. It also outperforms existing methods for PH detection of adults and to the best of our knowledge, this is the first work on multi-view video-based automated assessment of PH in newborns. Additionally, with our post-hoc explainability method, we are the first to show that a PH assessment model focuses on clinically relevant cardiac structures, aligning with how clinicians subjectively assess PH.

The thesis is structured as follows: In Chapter 2 we discuss pulmonary hypertension and review the existing methods for automatic PH detection. We also review methods for explainable ML in healthcare and for multi-modal learning, and explain methods for spatio-temporal classification. In Chapter 3 we describe our proposed method; that is the data processing, classification (for both the spatial and spatio-temporal case) and the post-hoc explainability approach. In Chapter 4 we report the results of the experiments performed to evaluate our different approaches and discuss the main findings. Finally, in Chapter 5, we conclude and discuss future work and limitations.

Chapter 2

Background

2.1 Pulmonary Hypertension in Newborns

Pulmonary hypertension (PH) is a rare but complex and progressive disease of the pulmonary arterioles, which can affect newborns, children and adults. Functional and anatomical changes result in an increase of pulmonary artery pressure (PAP), and PH is formally defined as mean PAP (mPAP) $>$ 25 mm Hg [81]. PH has unfavorable prognosis, and research consistently shows that prognosis is associated with the severity of the disease at diagnosis, thus, delayed treatment decreases the chance of survival [4, 81].

2.1.1 Human PH Estimation

The gold standard for PH diagnosis is measuring mPAP with right heart catheterization (RHC), but because this is an invasive and costly approach, it is primarily used to confirm PH diagnosis. Transthoracic echocardiography (ECHO) performed by experts is instead the recommended non-invasive diagnostic tool for estimating the likelihood of PH and the severity of PH [47]. ECHO consists of a sequence of ultrasound images of the beating heart from different angles of the heart (*views*), obtained from different locations of the transducer. Various different ECHO modes are available, including 2D-, 3D- and Doppler ECHO. Although 3D-ECHO is superior to 2D-ECHO, it can not always be performed and has not yet translated to routine clinical usage [27]. Thus PH evaluation is more commonly performed on 2D and Doppler ECHO, and mainly involves measuring ECHO variables (discussed in Section 2.1.1) and/or subjective ECHO evaluation (discussed in Section 2.1.1). However, these ECHO-based methods for PH estimation are time-consuming and expertise-demanding, which may delay care to a more advanced stage, potentially decreasing the chance of survival [4]. This raises the need for an automatic method to detect PH from ECHOs.

Quantitative Evaluation of ECHO

PH estimation frequently involves measuring various echocardiographic variables of 2D-ECHO and Doppler ECHO that allow for estimating mPAP. The traditional approach assumes the presence of tricuspid regurgitation (TR) in PH patients, and relies on measuring the TR velocity (TRV) from Doppler ECHO of apical four-chamber (A4C) and/or parasternal long axis views (PLAX). Systemic PAP (sPAP) can then be estimated from the TRV and right atrial pressure (RAP), as described in Equation 2.1. Importantly, RAP is not measured but estimated from inferior vena cava (IVC) diameter and inspiratory collapse. Finally, mPAP has a strong linear relationship with sPAP, and can be derived from sPAP with Equation 2.2 [7].

$$sPAP = 4 * (TRV)^2 + RAP \tag{2.1}$$

$$mPAP = 0.61 * sPAP + 2mmHg \tag{2.2}$$

Previous studies have demonstrated that the agreement between PAP estimated from TRV and invasively measured PAP is only moderate [11, 19, 23], and on the individual level significant under and over estimation can occur, possibly leading to misdiagnosis and inappropriate treatment [3]. There are number of reasons for this. First, as TRV is squared in Equation 2.1 even small errors in the absolute measurement of TRV can result in significant changes to the estimate of sPAP. Secondly, in many patients, IVC dimensions for RAP estimation cannot be obtained. Thirdly, absence of TR is insufficient to exclude the presence of PH - a recent study has for example shown that invasively confirmed PH is present in nearly half of patients without a reported TR who are also referred for RHC [48]. Measurements of further variables is thus recommended, especially in the absence of TR. These include variables measured from 2D ECHOs, such as the left atrial (LA) to aortic ratio (LA:Ao) from the PLAX view, which correlates with increased pulmonary flow [17]. Nevertheless, as no single variable has been detected as the ultimate predictive parameter to assess PH, and because the measurement may be frequently inaccurate, quantitative evaluation of 2D or Doppler ECHO is not the ultimate predictive tool, although widely used [19].

Subjective Evaluation of ECHOs

Since elevated PAP can result in abnormalities in the shape and structure of the heart, visual and subjective evaluation on 2D ECHOs is also commonly performed for estimating PH [17, 20]. The parasternal short axis view (PSAX) is specifically suitable for a subjective echocardiography evaluation of PH. From this view, abnormalities can for example be detected in the shape of the interventricular septum (IVS) and left ventricle during minimum expansion of the heart (systole). In a normal heart, the IVS is round, but becomes flat in patients with moderate PH, and in severe PH the left-ventricle becomes D-shaped, or crescentic, as seen in Figure 2.1. During maximum expansion (diastole), reversed volume of the ventricles can also be detected [17].

Other views, such as the parasternal long axis view (PLAX) and the apical four-chamber view (A4C), can also be utilized for subjective evaluation of PH. Changes in IVS shape can be seen from the PLAX view, and from the A4C view changes in the right-ventricular area are often detected in case of moderate and severe PH [17].

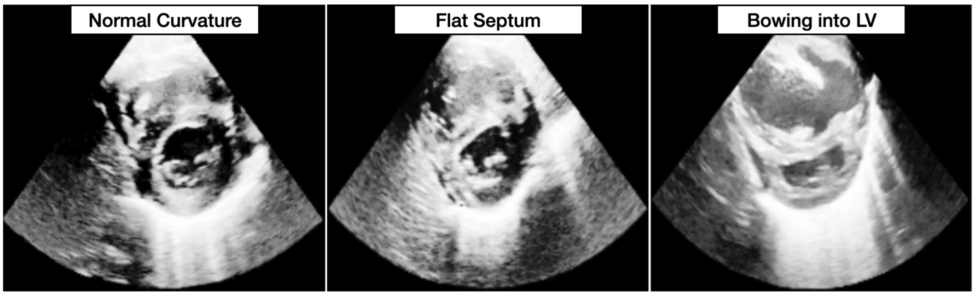


Figure 2.1: Varying septal morphology depending on the degree of PH on the short-axis parasternal view. Left: No PH, Middle: Moderate PH, Right: Severe PH. The examples are taken from our dataset.

2.2 Related Work

2.2.1 Machine Learning Approaches for PH Prediction

Several machine learning methods have been proposed to automatically estimate PH in adults using different input modalities, such as chest X-rays [80, 33], ECGs [34, 44], heart sounds recorded by acoustic sensors [28], CTs [71], and MRIs [13, 5]. However, not much effort has been directed towards the automatic assessment of PH using echocardiography, even though it is the recommended non-invasive modality for PH estimation, and the most common routine test used in newborns to diagnose or rule out various heart diseases [47].

The two exceptions are the work of Leha et al. [37] and Zhang et al. [75], which propose methods for automatic PH prediction in the adult population. The former approach relies on manually extracted ECHO parameters and applies various machine learning algorithms, such as regression and SVM, to these features, in order to predict PH. The goal of their approach is to help standardize and simplify integration of the several parameters that relate to PH. The main drawback is that the ECHO parameters must still be measured and estimated by highly trained specialists, so this approach does not help reduce the workload of experts. The latter approach [75] shows the potential of using deep learning for predicting PH using ECHOs, requiring no manual feature extraction. This method, however, has several limitations. First of all, it only uses a single view of the heart (A4C), although the literature has shown that considering multiple views improves accuracy for the manual prediction of PH [56]. Second, it works on static frames of the ECHO videos and does not exploit the spatio-temporal patterns in the ECHO sequence, although spatio-temporal deep learning methods have shown superior results for various video classification tasks, as further discussed in section 2.3. Finally, similar to the existing approaches for PH prediction from other modalities, this method has limited accountability and clinical usability. The reason is twofold; first, the black-box nature of these approaches makes their internal mechanisms and their results opaque, and second they focus only on binary PH classification but do not predict the PH severity. Severity estimation of PH is of greater clinical importance than PH detection, as guidelines for PH treatment type and urgency depend on PH severity, and the estimation of the severity of PH from ECHO is more challenging for cardiologists, with around 47% agreement to RHC, compared to 79% agreement for binary PH detection [12, 19].

2.2.2 Explainable Machine Learning in Healthcare

In recent years, interpretability and explainability of machine learning (ML) models have attracted much attention, and various methods aimed to help explain the reasons for a model's prediction have been proposed. This is especially true for the application of ML to healthcare, where achieving high predictive accuracy is often as important as understanding the prediction. Indeed, the lack of explainability is a key factor that limits wider adoption of ML in healthcare, as without it, medical practitioners often find it challenging to trust ML models [68].

Although explainability and interpretability of ML models are often used interchangeably, interpretability technically refers to the extent to which a model can be understood by a human on its own, whereas explainability refers to the extent to which the internal mechanics of a model can be (post-hoc) explained in human terms - usually for models which on their own are too complicated to be understood by humans [54]. In this thesis, we will focus on the explainability of ML methods. We specifically consider methods to explain the non-interpretable convolutional neural networks (CNNs), which are one of the most common deep learning methods for 2D and 3D medical image understanding, but their black-box nature limits clinical usability.

Explainability of CNNs can be achieved by using *visual explanation methods*, which identify and visualize the contribution of each pixel to the output of the trained network [43]. Generally, the results are expressed as an importance map (often referred to as salience map or attribution map) of the same size as the input image, where each scalar in the map quantifies the contribution of the corresponding pixel [39, 43]. The explanations are either generated by perturbing parts of the image and observing the change of the prediction (*perturbation-based methods*), or by computing the gradient of the prediction with respect to input features (*gradient-based methods*). As the gradient-based methods are faster to compute, they are more commonly used, although the perturbation-based methods have the benefit of not requiring access to the intermediate layers [43]. A vast number of gradient-based methods have been proposed, including Vanilla Gradients [60], DeconvNet [74] and Grad-CAM [57]. While they have been shown to explain model decisions [57, 35], some of those methods have also been shown to be insensitive to model and data, acting more like edge detectors by simply highlighting strong pixel changes in images. Of the tested methods, only Vanilla Gradients and Grad-CAM passed the insensitivity check, making them the preferred methods [1, 21].

In the field of medical imaging, the predictions of CNNs can be further explained by utilising expert- and domain-specific medical knowledge. For example, Zhu et al. [79] explain the predictions of lung diagnosis models by automatically generating anatomical features according to guideline criteria, and then using a perturbation-based method to calculate an importance map, showing the impact of each feature. Furthermore, Lee et al. [36] propose a deep network to explain the diagnostic decision of a malignant mass classifier with a visual pointing map and a diagnostic sentence justifying result simultaneously. Their proposed justification generator, which could be constructed on top of any malignant mass diagnosis network, is trained on medical reports for a given problem, and they use the task of diagnosis of breast masses to verify their approach.

The first application of interpretation frameworks to understand deep learning models from ECHOs has just recently been proposed [22]. Using visual explanation methods, they show that their models (trained on static ECHO frames) pay appropriate attention to key cardiac structures when performing human-explainable tasks, such as detecting the presence of pacemaker and defibrillator leads. However, no effort has yet been made to explain the predictions of models assessing PH, and no published work has shown spatio-temporal interpretation of ECHO sequences. Indeed, most work on explainable ML is centered around spatial input, although visual explanation approaches have recently been shown to be expandable to 3D CNNs trained on video clips [39]. For example, [67] and [66] recently adapted Class Activation Mapping (CAM) for 3D CNNs, such that important spatio-temporal regions of the input videos are highlighted, and Grad-CAM is inherently applicable to 3D-CNNs.

2.2.3 Multi-Modal Learning

Combining complementary information from multiple modalities is appealing for improving the robustness and performance of ML approaches, as signals from different modalities often carry complementary information. Various methods exist for combining the modalities in multi-modal learning, including early and late fusion [61], model ensemble [14], and joint end-to-end training methods where modalities are combined in the embedding space [45, 46, 41].

A few multi-modal learning methods have been proposed in the field of medical imaging, aiming to improve performance and/or mimic human workflow for tasks where clinicians usually consider multiple modalities. For example [53] combine (single-view) ECHOs and CMRs for predicting response to cardiac resynchronisation therapy, and [63] combine (single-view) ECHOs and ECGs for the prediction of hypertrophic cardiomyopathy. However, limited research has been directed towards integrating different modalities of ECHOs, i.e. the different views, although for most tasks, medical guidelines recommend the use of more than one view [56]. To the best of our knowledge, our work is the first to explore the benefits of combining multiple ECHO views in a deep learning setting.

2.3 Spatio-temporal Methods for Video Classification

Deep Learning models are able to capture not only spatial dependencies (e.g. CNNs) and temporal dependencies (e.g. RNNs), but also spatio-temporal dependencies (e.g. 3D-CNNs). Spatio-temporal deep learning methods have shown superior performance to spatial-only approaches for various video classification tasks, such as human action recognition [62, 72] and prediction of physical interaction forces [30]. Yet, this approach has not been widely explored in the context of medical video classification, partly because spatio-temporal models are more complex and thus more prone to overfitting on smaller datasets, which tends to be the case for labelled medical video datasets. However, following the recent release of a large medical video database EchoNet-Dynamics [50], a spatio-temporal benchmark approach for the dataset was proposed, showing improved results on various ECHO tasks compared to training on manually curated still images [51]. This included left-ventricle segmentation and estimation of ejection fraction, but the task of PH prediction could not be explored, as the dataset does not include PH annotations.

Spatio-temporal models inherit the complexity of both spatial and temporal models, but additionally they have the complexity of combining the two domains. As a result there is a great variety of spatio-temporal architectures, which can broadly be categorized into Convolutional Recurrent Neural Networks (CRNNs) [15] and 3D Convolutional Neural Networks (3D-CNNs) [52]. Below, we will discuss these two main categories.

2.3.1 Convolutional Recurrent Neural Networks

Convolutional Recurrent Neural Networks (CRNNs) are a family of models that combine CNNs with RNNs in one way or another. The two most common categories of CRNNs are Convolutional LSTM Networks (Conv-LSTMs) [59] and CNN-RNNs [16]. In ConvLSTMs, the general gate activation of an LSTM (i.e., the internal matrix multiplications) is replaced by the convolutional operation, allowing the data that flows through the cells to keep the input dimension (3D), and thus the network is able to exploit an extracted 3D tensor as the cell state [59]. In CNN-RNN networks, a CNN and an RNN are concatenated to formulate a collaborative network. The RNN is placed after a CNN, directly taking the output feature vector from the CNN as the input sequence, as shown in Figure 2.2(a). This network can then be jointly trained to learn temporal dynamics and convolutional representations. One of the benefits of the CRNN methods is that any type of RNNs can be used, such as LSTM, bidirectional LSTM or GRU, and attention can easily be added to the RNNs [16].

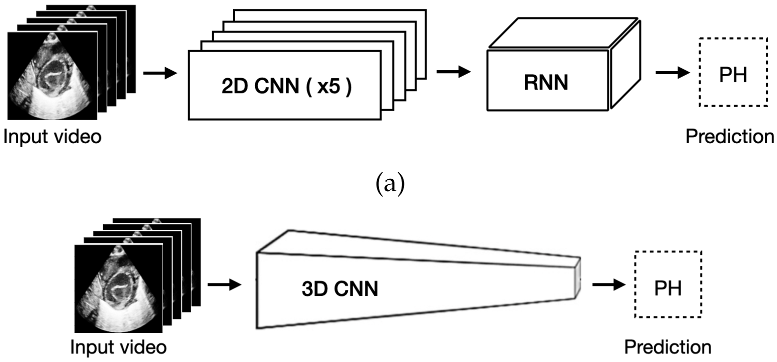
2.3.2 3D-CNNs

3D-CNNs are an extension of traditional CNNs, for spatio-temporal data. While CNNs use 2D kernels to perform 2D convolutions (which convolve images for extraction of spatial features), 3D-CNNs use 3D kernels to perform 3D convolutions of 3D cubes formed by stacking multiple video frames [2]. Thus, 3D-CNNs are homogenous networks that analyse spatial and temporal information in a single framework, directly extracting spatio-temporal features. This is different from the heterogeneous CNN-RNNs, where the CNN analyzes the spatial information, and the RNN concurrently manages temporal information [30]. Figure 2.2 summarises the architectural differences of CNN-RNN and 3D-CNN models.

The 3D CNN has an inherent disadvantage of high computational complexity and memory usage due to a large number of parameters P, as given by the following formula:

$$P = N \times C \times T \times (W \times H \times 1) \tag{2.3}$$

where N is the number of kernels/filters, C is the number of channels, T is the number of stacked frames, and (W, H) is the spatial size of the kernel. The T parameter is additional compared to 2D CNNs, and thus the computational complexity of 3D-CNNs is increased according to the number of sequential frames used as input. As a result of the large number of parameters, 3D-CNNs tend to overfit on small datasets.



(a)

(b)

Figure 2.2: (a) Heterogeneous network structure in the CNN-RNN method and (b) homogeneous network structure in 3D CNN method, where the input video consists of 5 adjacent frames.

Multiple variations of 3D-CNN architectures have been proposed, and different categorisations of these architectures are possible. We propose grouping the 3D-CNN architectures into three categories based on the integration of spatio-temporal convolutions:

- **Full 3D CNNs:** This group entails the traditional 3D-CNN architectures which consider the spatial and temporal dimensions jointly in all convolutional layers. Variations within this category differ mainly in the backbone network. Networks in this category include the *C3D* [69], which does not exploit any pre-trained 2D networks, the *I3D* [9], with pre-trained 2D Inception-V1 as a backbone, and *ResNet3D* [25], with pre-trained 2D ResNet as a backbone.

- **Partial 2D/3D CNNs:** This category consists of models that provide a middle ground between two-dimensional convolutions that consider only spatial relationships and the three-dimensional convolutions. They either factorize the 3D convolutional filters into separate 2D spatial and 1D temporal components, like *S3D* [73] and *R(2+1)D* [70], or employ spatio-temporal convolutions only in the early layers, but spatial convolutions in the top layers, such as the *MCResNet* [70]. These networks have been shown to allow for easier optimization, often resulting in improved performance compared to full 3D-CNNs [70].

- **Multiple 3D-CNN pathways:** These are models that employ two parallel 3D-CNN pathways, a pathway operating on a slow frame rate (i.e. with a large temporal stride) to capture spatial semantics, and a pathway operating on a high frame rate (i.e. with a small temporal stride) to capture motion at a fine temporal resolution [18]. The *SlowFast* [18] network was the first proposed model of this sort, but subsequently a few variations of it have been proposed [38].

In this thesis, we will employ 3D-CNNs rather than CRNNs, because learning spatio-temporal features simultaneously from videos is generally more effective than learning them separately [78, 30]. Furthermore, although all spatio-temporal methods are in inherently computationally heavy, analysing the spatial and temporal information in a single framework is advantageous for weight parameter reduction [30]. We specifically compare architectures from the three 3D-CNN categories described above.

Chapter 3

Methods

We propose two methods for PH assessment of newborns, namely a *spatial approach*, training a CNN on ECHO frames, and a *spatio-temporal approach*, training a 3D-CNN on ECHO sequences. A generic overview, relevant for either the spatial or the spatio-temporal method, is shown in Figure 3.1(a), when training on a single view. To increase the robustness of our method, we propose extending it to using multiple views. We employ different *multi-view approaches*, including majority voting, visualised in Figure 3.1(b). Finally, to increase the accountability of our method, we complement our predictions with saliency maps from each view, as seen in Figure 3.1(c).

In this chapter, we explain our proposed method in more detail. In Section 3.1 we describe the dataset at hand and the data processing and augmentation performed, as well as methods to tackle the data imbalance. In Sections 3.2 and 3.3 we explain the details of the single-view spatial and spatio-temporal approaches, respectively. The multi-view method is then discussed in Section 3.4, where we describe on one hand an ensemble approach (majority voting) for combining views, and on the other hand feature-level fusion. Finally we describe the explainability method in Section 3.5.

3.1 Dataset

The dataset used in this work consists of 2D transthoracic echocardiography videos (ECHOs) of 199 newborns from five different views, where each view shows the heart from a specific angle. The spatial size of the ECHOs is 1440×866, and the mean temporal size is 122 frames. As the ECHOs operate on 25 frames per second, the average video length is around 5 seconds, which accounts for around 10 heartbeats. The mean age of the newborns is 56 days and the mean weight is 2.9 kg. The five views include a parasternal long-axis view (PLAX), apical four-chamber view (A4C), and three parasternal short-axis views; at the level of papillary muscles (PSAX-P), at the level of

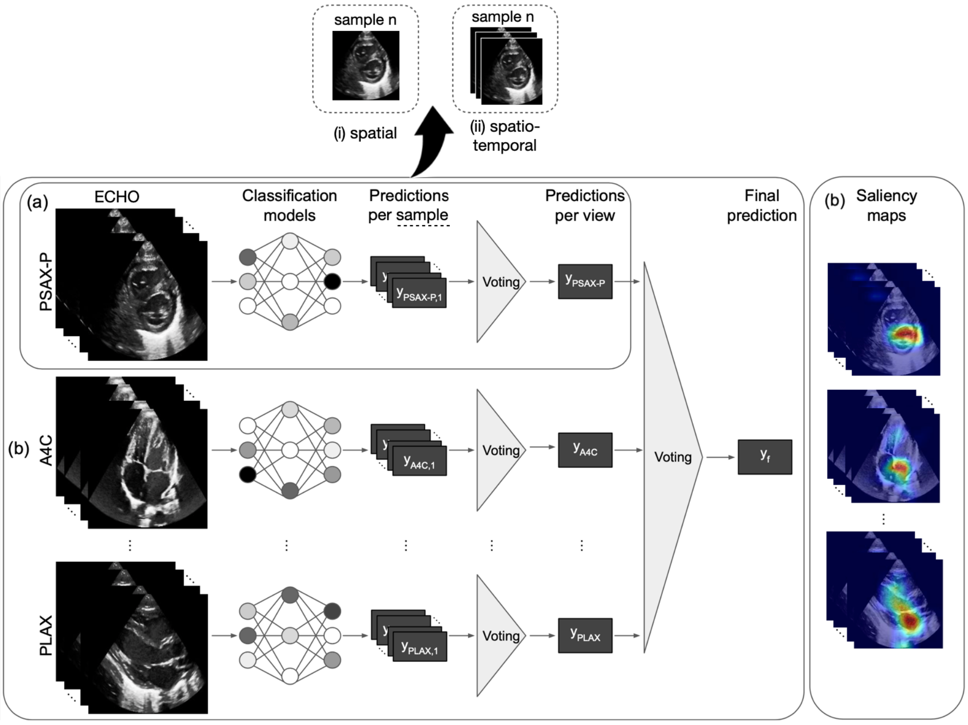
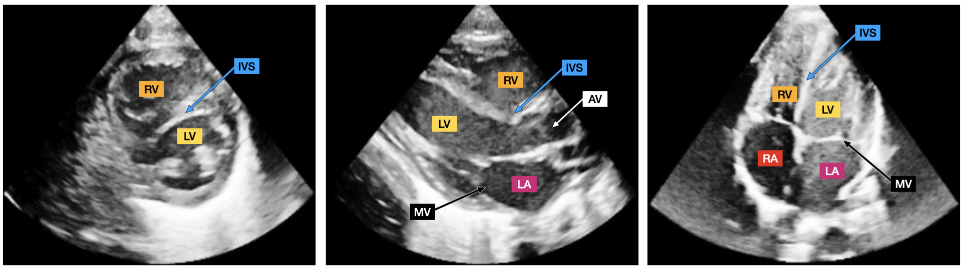


Figure 3.1: Overview of our proposed method to automatically assess PH using (a) a single view and (b) multi-view approach with majority voting. Both approaches are suitable for training on (i) spatial input or (ii) spatio-temporal input. Saliency maps (c) are provided from each view.

semilunar valves (PSAX-S), and on the apical short-axis (PSAX-A). Figure 3.2 shows an example of ECHO frames from three different views, one from each axis.

All data is pseudonymized, and approval for reuse of the data for our research was obtained from the responsible ethics committees, with written and oral consent obtained for each patient. It is collected by the Hospital Barmherzige Brüder Regensburg using GE Logic S8 ultrasound machine with the S4-10 transducer between the years 2019 − 2020.

The PH annotations, provided by a pediatric cardiologist, differentiate between *none*, *mild*, *moderate* and *severe* PH, as described in Table 3.1. Furthermore, a few samples have ambiguous annotations, such as between none and mild (*none-mild*), and between moderate and severe (*moderate-severe*). We explore different methods to categorise the ambiguous labels, which results in slightly different dataset sizes (see Section 3.1.1). The original dataset is heavily imbalanced, as only 68 subjects (34%) show some signs of PH, but



(a) PSAX-P (b) PLAX (c) A4C

Figure 3.2: ECHO frames for (a) PSAX-P view, (b) PLAX view, and (c) A4C view. Important cardiac features for each view have been labelled: LV = left ventricle, RV = right ventricle, LA = left atrium, RA = right atrium, IVS = interventricular septum, MV = mitral valve, AV = atrial valve.

the majority, or 131 subjects (66%), have no PH. Only 2 subjects ($<$ 1%) have severe PH. Additional details of the dataset is found in the Appendix (A.2).

Label	**Description**
No PH	No PH visible
Mild PH	Sub-systemic
Moderate PH	Approx. system pressure
Severe PH	High system pressure

Table 3.1: Description of PH annotations, as provided by experts.

3.1.1 Data Pre-Processing

The first data pre-processing step involves cropping and masking the ECHOs to eliminate information such as additional text or signals outside the scanning sector. We then resize the spatial size of the ECHOs to 224 × 224 pixels, using bilinear interpolation. This size was chosen to match the input size of the pre-trained models on the Imagenet [55] and Kinetics [29] datasets. Next, we improve the contrast of the frames by applying histogram equalization to each frame in an ECHO, thus distributing the pixel intensities to the full range of gray-scale values. We used the following histogram equalisation algorithm, defined in the OpenCV library [6]:

```
1: Calculate the histogram $H$ for the source image $src$
2: Normalise the histogram so the sum of the histogram bins is 255
3: Compute the integral of the histogram: $H^{'}_{f} = \sum_{0 \leq j \leq i} H(j)$
4: Transform the image using H' as a look-up table: $dst(x,y) = H'(src(x,y))$
```

Finally, we normalize the frames by dividing by the max pixel value (255.).

Apart from processing the ECHOs themselves, we also explore different approaches to generate training labels from the provided annotations.

For the task of binary PH detection, we binarize the annotations by distinguishing between subjects with and without PH. We eliminate the two subjects with severe PH, as these might be seen as outliers. We furthermore remove the five subjects with the label none-mild, as it is not evident whether these have a sign of PH or not. For the PSAX-P view, which is our baseline view, this results in a PH detection dataset of 192 subjects, and the resulting label distribution is described in Table 3.2. Note that slight variations are possible for the different views, as for a few subjects not all views are present.

For PH severity prediction, we define three PH severity categories; *no PH* (subjects with no PH), *mild PH* (subjects with original labels mild or mild-moderate), and *significant PH* (subjects with original labels moderate, moderate-severe or severe). Similar to the binary PH detection, we eliminate subjects with the label none-mild. However, as opposed to the binary classification, we include the two severe PH cases with the significant PH category, due to the higher data imbalance. This results in a PH severity prediction dataset of 194 subjects for the PSAX-P view. Table 3.3 describes how the labels are created, as well as the distribution of labels. By comparing Tables 3.2 and 3.3, we see that class imbalance is higher for the PH severity prediction than for the binary PH detection, with the minority class being only 16% of the dataset for the severity prediction, as opposed to 34% for the binary detection.

PH Label	Original Labels	Count	Percentage
No PH	No PH	126	66%
PH	Mild PH Mild - Moderate PH Moderate PH Moderate - Severe PH	66	34%

Table 3.2: Derived PH labels for the PH binary detection dataset of 192 subjects, with information on the size and ratio of each class.

3.1.2 Data Augmentation

In order to reduce the risk of overfitting, and to improve generalisation, we employ data augmentation during training, where each training sample has a 90% chance to be transformed, as determined empirically. We both implement *intensity transforms*, so that the learned model is invariant to intensity variations, as well as *spatial transforms*, to increase resilience towards perturbations that may occur due to different zoom settings of the US machine and/or different placements of the transducer [42]. Figure 3.3 shows few examples of the augmentations.

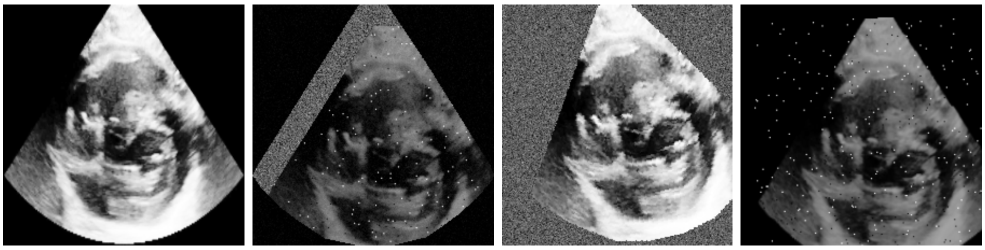


Figure 3.3: Examples of different augmentations of the same frame.

We employ the following six intensity transforms, each applied with a probability of 0.5 for a given sample:

1. random sharpness adjustments, sharpening the image by up to 8.0×, or blurring it with a sharpness factor (f) < 1.0, specifically $f \in [0, 0.9999]$.
2. random brightness adjustments, using an enhancement factor between 0.5 and 1.2, where a factor of 0.0 gives a black image, a factor of 1.0 gives the original image, and a factor > 1.0 gives a brighter image.
3. random gamma correction, with a gamma factor (γ) between 0.25 and 2.0, where the gamma correction (I_γ) of an input image I_{in} is given by: $I_\gamma = 255 * (I_{in}/255)^\gamma$
4. addition of salt and pepper noise, with a threshold of 0.005.
5. addition of random Gaussian noise.
6. random background colour variations, varying background colour between black and grey and with varying amount of speckle noise [8].

PH Label	**Original Labels**	**Count**	**Percentage**
No PH	No PH	126	65%
Mild PH	Mild PH Mild - Moderate PH	32	16%
Significant PH	Moderate PH Moderate - Severe PH Severe PH	36	19%

Table 3.3: Derived PH labels for the PH detection dataset of 194 subjects, with information on the size and ratio of each class.

We employ the following three positional transforms, each applied with a probability of 0.6 for a given sample:

1. random rotation of up to 15°
2. random translation of up to 0.1×
3. random rescaling, scaling down to 0.8× or zooming up to 1.2×

3.1.3 Class Imbalance Intervention

We explore two different methods to deal with class imbalance; sample weighting of the loss function and re-sampling the dataset, both described below. As initial experiments suggested better results with the prior, for all further experiments we use re-sampling to balance the data. Additionally, for the evaluation of our method, we include performance metrics that are suitable for imbalanced datasets, such as balanced accuracy (see Section 4.1).

Sample Weighting in the Loss Function

By introducing sample weighting in the loss function, we can impose an additional cost on the model for making mistakes on the minority class(es) during training. These penalties can bias the model to pay more attention to the minority class(es). We calculate the sample weights as the inverse of the class frequency and normalise them over the number of classes and samples, following the balanced heuristic in [31]. Equation 3.1 explains formally how the class weights m_c of each class c are calculated, as well as the sample weights w_i (which are assigned to the weights of the class they belong to)

$$\begin{aligned} m_c &= \frac{N}{(C * N_c)} \\ w_i &= m_{y_i} \end{aligned} \tag{3.1}$$

where N_c is the number of samples belonging to class c, N is the total number of samples, C is the number of classes, and y_i is the class label of sample i.

The weighted loss function L^W can then be defined as a function of any given loss function L_i, applied on sample i, and the sample weights w_i, as defined with the formula below:

$$L^W = -\sum_{i=1}^{N} w_i * L_i \tag{3.2}$$

Resampling the Dataset

By resampling the dataset, it is possible to simulate a more balanced dataset. For this, there are two general approaches:

1. *Oversampling*: Adding copies of samples from the under-represented class during training or as a pre-processing step.
2. *Undersampling*: Removing samples from the over-represented class during training or as a pre-processing step.

We use a combination of both approaches, specifically we apply random resampling with replacement, which has proven to be robust [40]. We perform the resampling on the fly during training, in combination with data augmentation, such that for each epoch, the model processes approximately equal number of samples from each class. To achieve this, we employ a weighted random sampler, which samples elements from the dataset with replacement from indices $i = [0, \ldots, N]$, with probabilities given by the weights $W = [w_0, \ldots, w_N]$, where each sample weight w_i is calculated according to Equation 3.1. Thus, for each epoch, some samples from the minority class(es) may be randomly selected more than once, and some samples from the majority class might not be selected. Note that the oversampling procedure will not produce identical copies of the same samples, because the data augmentation slightly modifies each sample.

3.2 Spatial Approach

For the spatial approach towards PH prediction, we train a convolutional neural network on manually curated still ECHO frames, extracting only spatial features from the videos. To overcome the scarcity of the annotated data, we extract n frames from each ECHO, using different frame extraction heuristics, as further explained in Subsection 3.2. Each frame is thus considered an individual sample for the classification, giving rise to frame-level predictions, whose results are aggregated to achieve view-level predictions, as explained in Subsection 3.2.1. The classification training details and the model architecture is further described in that subsection. An overview of the spatial approach is provided in Figure 3.4, both for the (a) frame extraction phase, and (b) classification phase.

As a baseline approach for extracting training frames from each ECHO, we simply select n random frames per ECHO. Additionally, we are interested in extracting the frames that are most relevant for PH assessment performed by cardiologists, that is, the frames corresponding to minimum- and maximum expansion of the heart We explore two different approaches to identify the minimum- and maximum-expansion frames; an algorithmic approach and a

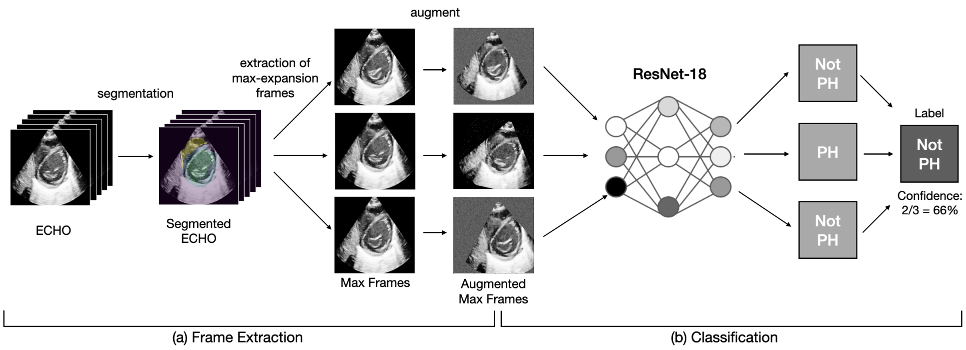


Figure 3.4: Overview of the spatial method for PH detection using a single view (PSAXP-P), when training on three maximum-expansion frames per ECHO (and for simplicity, only for a single patient). The first step involves (a) extraction of maximum-expansion frames using segmentation. The next step involves (b) classification using the extracted frames.

segmentation approach, but as explained below the segmentation approach is the one used for further experiments.

Algorithmic Approach

By making use of an existing algorithmic approach for cardiac phase detection [76], we can identify the frames corresponding to systole (i.e. minimum expansion) and diastole (i.e. maximum expansion). A drawback of this method is the large amount of parameters that need to be tuned and set according to population-specific knowledge. Furthermore, the generalisability of the method to newborns is unclear. After setting the most important parameters according to average newborn cardiac statistics, and running the algorithm on our dataset, results were promising but not sufficient. Since off-by-one errors in the heart-phase estimation of the first cycle escalate in proceeding cycles, it is not suitable for identification of all the minimum- or maximum expansion frames of an ECHO consisting of multiple cycles.

Segmentation Approach

Training a segmentation model on ECHO frames to identify pixels corresponding to the left- and right- ventricles, allows for calculating for each frame the relative area of the ventricles. This is simply done by dividing the pixel count corresponding to the ventricles by the total number of pixels. The minimum- and maximum-expansion frames are then easily identified as the frames with the smallest or largest relative ventricle area, respectively. However since our data does not contain annotations of the ventricles, we are unable to train a segmentation model on our dataset. Instead, we make use of publicly available ECHO segmentation models [75] that have been trained on adult ECHOs across five common views. Three of these views are the same views as in our dataset; PSAX-P, PLAX and A4C. The intersection over union (IoU) score on their own test data for the views and segmentation areas of interest ranges between 64.6 and 88.9, as seen in Table 3.4. Inconveniently, the poorest segmentation performance is for the PSAX-P view, which is our baseline view. Additionally, we expect that applying these models to our dataset will result in slightly lower performance, as the segmentation models are trained on a dataset with mean subject age of 59 [75], whereas our dataset population is newborns. This assumption can not be verified numerically, due to lack of ground truth for our data, but visual inspection suggests this is the case. As the segmentation approach is more robust than the algorithmic approach, it is our method of choice for all experiments involving minimum- or maximum-expansion frames. Also note that since we are only interested in the joint area of the left- and right-ventricles relative to the rest of the heart, perfect segmentation of each ventricle is not required. The segmentation approach for extracting three maximum-expansion frames per ECHO is visualised in Figure 3.4(a).

View	No. Training Frames	Segmented Area	IoU Accuracy
PSAX	124	Left ventricle	79.6
		Right ventricle	64.6
PLAX	130	Left ventricle	87.9
		Right ventricle	85.2
A4C	182	Left ventricle	88.9
		Right ventricle	83.3

Table 3.4: Performance of SOTA ECHO segmentation models [75], as well as the training data size for each model.

3.2.1 Classification

For the PH classification we use a ResNet-18 [26] consisting of convolutional layers with residual connections connecting odd-numbered layers, as seen in Figure 3.5. We initialise the model weights with pre-trained weights from Imagenet [55], and train it to minimize the cross-entropy loss between the prediction and true class. When exploring with weighted cross-entropy loss, we use Equation 3.2 in combination with the sample-wise cross-entropy loss. We employ a batch size of 64 and train for up to 300 epochs (or until early stopping), using Adam optimizer with a learning rate of 0.001.

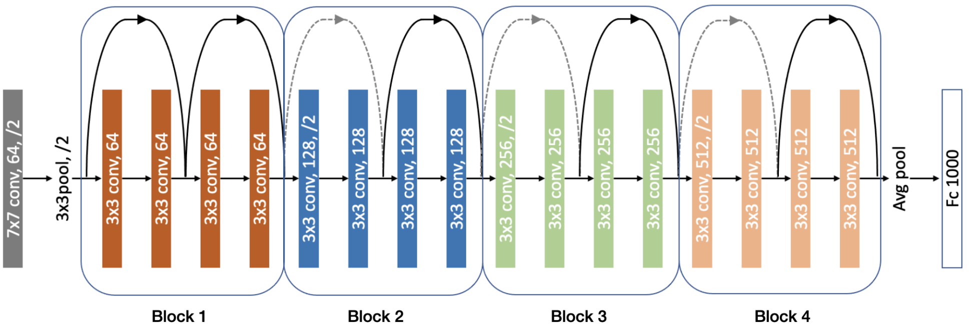


Figure 3.5: ResNet-18 architecture, consisting of 18 layers, of which 17 are convolutional layers, and one is a fully connected layer. The dotted lines represent residual skip connections.

The spatial model is trained on n frames per ECHO, and thus it provides a prediction for each frame. To get view-level results, we aggregate frame-level predictions of a given view $\{y_{\text{view},i}\}_{i=1,\ldots,n}$ through majority voting, i.e. by selecting the most frequently predicted label as the view-level prediction y_{view}. The view-level confidence is then defined as $C = |y^*_{\text{view}}|/n$, where $|y^*_{\text{view}}|$ is the count of the most frequently predicted label from the list of predictions for the n frames of a given ECHO per view. Figure 3.4(b) summarises the classification process when training on maximum-expansion frames after the maximum-expansion frames have been extracted and augmented. Note that the process looks similar for random frames, or minimum-expansion frames.

3.3 Spatio-Temporal Approach

For the spatio-temporal approach, we integrate spatial as well as temporal information into the learning process, thus, training on sequences instead of single frames. This mitigates the frame-level variations that can occur due to external changes, such as the position or the contact of the transducer, or in the cardiac function itself. Furthermore, it eliminates the need for using segmentation to extract frames corresponding to systole and diastole, given that the training sequences cover on average at least one heartbeat. An overview of the single-view spatio-temporal approach is provided in Figure 3.6. Similar to the spatial approach, n samples are extracted from each ECHO for training, but in this case, each sample is a shorter video sequence. Subsection 3.3.1 explains further how these sequences are extracted. These sequences are then the input to a spatio-temporal model, as described in Subsection 3.3.2.

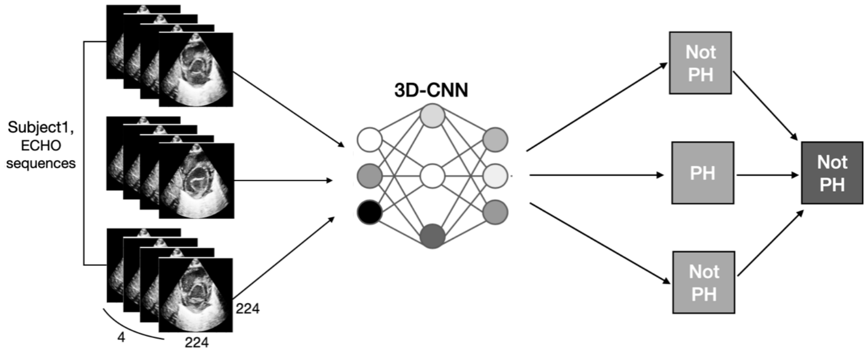


Figure 3.6: Overview of the single-view (PSAX-P) spatio-temporal classification process, when using $n = 3$ sequences per ECHO, each of length $k = 4$ frames. A single input sample thus is a cube with the shape $224 \times 224 \times 4$.

3.3.1 Extraction of Sequences

From each ECHO of length t (i.e. ultrasound video consisting of t frames), we extract n random sequences consisting of $k \leq t$ frames. Specifically, each sequence is extracted by randomly choosing a frame as the starting frame, followed by selecting every s_{th} frame, until total k frames have been selected. For a sampling interval $s = 1$, this involves selecting k consecutive frames. Given the sequence length k and sampling interval s, the effective length l is defined as $l = k * s$, and it determines the number of frames the given sequence covers. The effective length can in theory be set to any number between 1 and max frames, where max is the length of the full ECHO, on average $max = 122$ frames. However, a sequence covering a single frame does not utilize the temporal information, and is thus not useful. Furthermore, using an entire ECHO as a sequence leads to slow training and only provides for a single sequence per ECHO, resulting in fewer training samples. We suspect a sequence covering at least a full heartbeat (i.e. $10 - 12$ frames) will be necessary for best performance, but the ideal effective length, sequence length and sampling interval is determined with an ablation study. Note that the same augmentation is applied to each frame in a given input sequence.

3.3.2 Classification

For the spatio-temporal approach we employ different variations of 3D-CNN models, all having in common residual connections and spatio-temporal convolutions across frames, as well as a softmax activation function.

The output of each layer, that is the value of the j_{th} feature map (v) in the i_{th} layer at position (x, y, z), is formally described in the following formula:

$$v_{i,j}^{x,y,z} = \text{softmax}\left(b_{ij} + \sum_{m} \sum_{w=0}^{W_i-1} \sum_{h=0}^{H_i-1} \sum_{t=0}^{T_i-1} \alpha_{ijm}^{wht} v_{(i-1)m}^{(x+w)(y+h)(z+t)} \right) \tag{3.3}$$

where W_i and H_i represent the spatial size (width and height) of the 3D kernel in the i_{th} layer, and T_i the temporal size of the kernel. α_{ijm}^{wht} is the $(w, h, t)_{th}$ weight of the kernel connected to the m_{th} feature map in the previous layer.

We implement three 3D-CNN architectures with variable integration of temporal convolutions, summarised in Table 3.5. The first one is an 18-layer ResNet3D [25], visualised in Figure 3.7(a). It is a full 3D-CNN using solely the vanilla spatio-temporal convolutions described in Equation 3.3. The second one is an 18-layer R(2+1)D [70], a partial 2D/3D network that factorizes the 3D convolutional filters into separate 2D spatial and 1D temporal components (see Figure 3.7(b)). The last one is a 50-layer SlowFast model [18] which employs two parallel 3D-CNN pathways, a slow pathway with a large temporal stride to capture spatial semantics, and a fast pathway with a small temporal stride to capture motion at fine temporal resolution (see Figure 3.7(c)). Note that we have selected one architecture from each of the three 3D-CNN categories described in Section 2.3.2, for maximum diversity. All architectures are initialised with pre-trained weights from the Kinetics-400 dataset [29]. We employ a ResNet-18 backbone for the first two networks, whereas the last one is implemented with ResNet-50 backbone, since pre-trained weights are only available for SlowFast with ResNet-50 backbone. Figure 3.7 visualises the structural differences of the three models.

Model	**Category**	**Backbone**
ResNet3D	Full 3D CNN	ResNet-18
R(2+1)D	Partial 2D/3D	ResNet-18
SlowFast	Multiple 3D-CNN Pathways	ResNet-50

Table 3.5: The spatio-temporal architectures, their category and backbone.

The spatio-temporal models are trained on n random sequences of length k from each ECHO, and view-level results are achieved by aggregating results through majority voting from all sequences for a given ECHO, in a similar manner as for the spatial approach, explained in Section 3.2.1. The models are trained to minimize the cross-entropy loss using an ADAM optimizer with a learning rate of 0.001, and a batch size of 8. The smaller batch size compared to the spatial approach is required, due to increased memory requirements of 3D-CNNs. The models are trained for up to 300 epochs, or until early stopping.

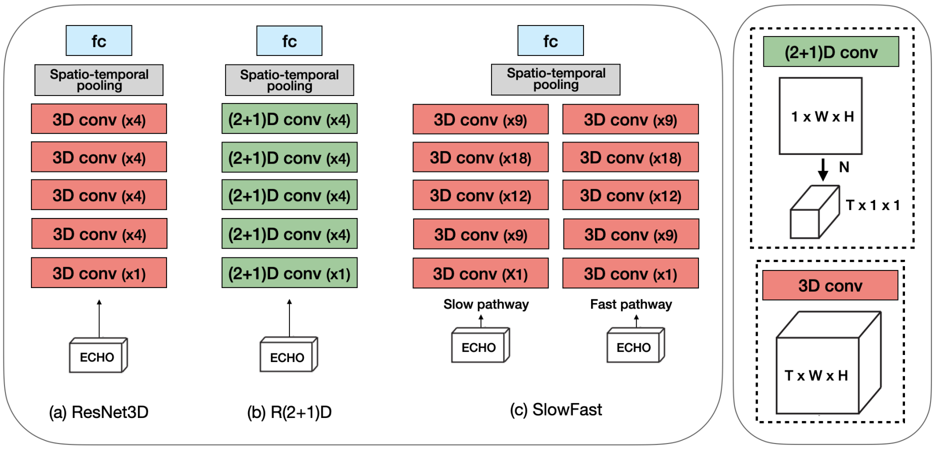


Figure 3.7: To the left, we show a high-level structure of the three spatio-temporal architectures used: (a) 18 layer ResNet3D, (b) 18 layer R(2+1)D (b) and (c) 50 layer SlowFast. The difference between the 3D conv and (2+1)D conv blocks is visualised to the right. T is the temporal dimension, and (W, H) the spatial dimensions. N is the number of 2D filters in the (2+1)D conv block.

3.4 Multi-View Approach

To increase the robustness and performance of the method further, we employ a multi-view approach, which can be used as an extension to both the spatial and spatio-temporal approach. Note that we use view annotations to differentiate the distinct views, but following recent work on view classification [75], our method can easily be extended to incorporate ECHOs without annotation. We explore two different approaches for view aggregation, an ensemble approach with majority voting and a feature-level fusion. Figure 3.1(a) shows the multi-view approach when using majority voting.

3.4.1 Ensemble

For the ensemble approach towards multi-view classification, we train a separate model for each available view, and achieve a final prediction, y_f by majority voting of the individual view-level predictions. In the case of a tie, the prediction of the model(s) with higher confidence is selected. The benefit of this approach is that given trained models on individual views, the extension to a multi-view prediction is trivial, and requires no further training. Furthermore, ensemble methods provide a way to potentially reduce the variance of the predictions, which can result in a better average performance as compared to any single model [49].

3.4.2 Feature-Level Fusion

This approach involves combining the modalities in the embedding space by learning intermediate features for each view, which are then combined and jointly modeled to make a decision. Specifically, for each view, a separate feature extractor is constructed, and the learned, intermediate features are combined via concatenation or summation. The combined features are then input into a final classification layer which provides the subject-level prediction y_{f}. The benefit of this method is that a single end-to-end model is trained on learned features from all available views, such that only one model has to be trained, instead of one for each view. The drawback is that it is more memory-heavy, especially when used with a spatio-temporal model.

3.5 Explainability

To increase the accountability and clinical usability of our method, we complement our predictions with saliency maps for each view, providing an interpretable explanation that mimics the clinical workflow. For the spatial approach, we provide spatial saliency maps that highlight important pixels, and for the spatio-temporal approach, we provide spatio-temporal saliency maps highlighting important spatio-temporal regions, as in Figure 3.1(b).

Among different methods [64, 77, 57], we chose to use Gradient-weighted Class Activation Mapping (Grad-CAM), which has originally been proposed for 2D-CNNs trained on images. Grad-CAM exploits the gradients of any target concept flowing into any given convolutional layer, to produce a coarse localization map highlighting the important pixels in the input image for the decision of interest [57]. In order to provide explanations for the spatio-temporal model, we extend Grad-CAM to 3D-CNNs processing spatio-temporal video inputs, allowing us to identify the spatio-temporal regions on the video sequence that the network finds most informative for its prediction. We do this by assigning each neuron a relevance score for the class prediction at the output layer, and backpropagate this information to the last convolutional layer to produce a coarse spatio-temporal localization map highlighting spatio-temporal areas of the ECHO video sequence. Specifically, let $A_1, \ldots, A_k$ be the k spatio-temporal feature maps in the last convolutional layer. We then decide the importance of each of the k feature maps for the class of interest, c, by weighting each item of each feature map with the gradient, and get a 3D heatmap that highlights spatio-temporal regions that positively or negatively affect the class of interest. This heatmap is sent through a ReLU function, to remove all negative values, as we are only interested in the parts that contribute to the selected class c. We finally scale the map for visualization purposes and overlay it over the original image.

Formally, the localization map L^c for class c, before scaling for visualization, is defined with the formula:

$$L^c = ReLU(\sum_k \alpha_k^c A^k) \tag{3.4}$$

where α_k^c are the importance weights, obtained by global average-pooling of the gradient of y_c (the score of class c) with respect to A:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \sum_z \frac{dy_c}{dA_{ijz}^k} \tag{3.5}$$

Chapter 4

Results and Discussions

In this chapter we will report and discuss the main results of our method for both PH detection and PH severity prediction. After describing the setup of the experiments in Section 4.1, we report the results of the spatial approach in Section 4.2 and the spatio-temporal approach in Section 4.3. Finally, we report the results from the explainability analysis in Section 4.4.

4.1 Experimental Setup

Experiments were run on a cluster containing different NVIDIA GeForce graphic cards (see Appendix A.1). Spatial models are trained with a single GPU, but spatio-temporal models with two GPUs. For all experiments, a stratified 10-fold cross-validation with replacement was performed (also known as out-of-bootstrap estimate), such that the data was randomly split 10 times into 20% validation set and 80% training set. Note that the splitting into training and validation sets was done on a patient basis. As classification metrics, we evaluated the balanced accuracy, weighted F1-score, weighted precision, weighted recall, and area under the receiver operation characteristic curve (AUROC). For the multi-class case, AUROC was calculated by comparing every unique pairwise combination of classes, i.e. using a one-vs-one scheme. Although AUROC is not a suitable metric for ensembles, for the sake of completeness we report it [in brackets] for our ensemble multi-view approach, calculating it based on the output probabilities of the most confident model selected by the ensemble. Results were averaged over the folds, and the mean and standard deviation are reported, as well as the average model confidence (as defined in Section 3.2.1). All results are reported per subject, but sample-level results are defined in Section A.4.1 in the Appendix.

4.2 Spatial Approach

In this section, we report the results of the spatial ResNet-18 model, as described in Section 3.2, for the detection and severity prediction of PH. Specifically, in Subsection 4.2.1 we discuss the ablation studies performed on the base view (PSAX-P) for the binary PH detection task. In Subsection 4.2.2 we provide an empirical assessment of the best spatial approach, as per the ablation analysis, for both the PH detection and the severity prediction, for all five views. We further report the results of the multi-view approaches.

4.2.1 PH Detection on the PSAX-P View - Ablation

To understand the importance of different regularisation methods, and the effects of different frame extraction methods, ablation studies were performed on the PSAX-P view, for the binary PH detection task.

Augmentation & Regularisation

We report in Table 4.1 the results of different regularisation techniques when applied to the spatial PSAX-P model for PH detection; specifically the following: Augmentation (*aug*), weight decay of 0.001 (*wd*), and initialising the model with pre-trained weights (*pre-trained*). The weight decay value of 0.001 was set empirically, and for this study we keep the number of frames per ECHO fixed, extracting 10 random frames from each ECHO.

The results clearly show the importance of regularisation, as using all three regularisation techniques gives an improvement of over 25% compared to using no regularisation. For example, F1-score improves from 0.72 when evaluating a baseline model with no regularisation, to 0.91 when evaluating a model trained with all three regularisation techniques. This verifies our assumption that regularisation is important to prevent overfitting, which is otherwise hard to avoid when training on a relatively small dataset like ours. Although the best results are achieved when using all three regularisation approaches, the data augmentation has the largest positive effect on performance. For all following experiments, models are trained with all regularisations, and with re-sampling the dataset.

Extraction of frames

Next, we study the effect of training on various number of frames per ECHO, using different frame extraction methods. In Table 4.2 we report the results of training a spatial PH detection model using different percentiles of maximum-expansion frames (*Max*) and minimum-expansion frames (*Min*) from the PSAX-P view. Note that the $\mathrm{P_{th}}$ percentile of maximum- or minimum-expansion frames corresponds to selecting frames that have left- and right-ventricle area that is larger or smaller than P% of the other ECHO frames for a given patient. As a comparison, we also report the results from training on 10 random frames (*Rand*-10), and all frames (*Rand*-all).

In general, selecting the frames corresponding to the minimum-expansion of the heart performs significantly better than selecting frames corresponding to the maximum-expansion of the heart. As an example, when selecting the top 95_{th} minimum-expansion frames (i.e. only around 7 frames per ECHO),

	Regularisation	AUROC	F1-Score	Precision	Recall	Balanced Accuracy	Confidence
Rand Weights	No aug, no wd	0.76 ±0.07	0.72 ±0.04	0.77 ±0.05	0.72 ±0.05	0.72 ±0.05	0.85 ±0.02
	No aug, wd	0.79 ±0.09	0.73 ±0.11	0.80 ±0.04	0.73 ±0.10	0.74 ±0.07	0.85 ±0.03
	Aug, no wd	0.91 ±0.04	0.89 ±0.05	0.91 ±0.04	0.89 ±0.05	0.89 ±0.05	0.85 ±0.02
	Aug, wd	0.91 ±0.03	0.89 ±0.04	0.90 ±0.03	0.89 ±0.04	0.89 ±0.04	0.86 ±0.02
Pre-trained	No aug, no wd	0.86 ±0.04	0.82 ±0.05	0.84 ±0.05	0.82 ±0.05	0.81 ±0.05	0.87 ±0.02
	No aug, wd	0.86 ±0.07	0.83 ±0.07	0.84 ±0.06	0.83 ±0.07	0.82 ±0.06	0.87 ±0.02
	Aug, no wd	0.92 ±0.05	0.90 ±0.04	0.92 ±0.03	0.90 ±0.04	0.91 ±0.04	0.86 ±0.03
	Aug, wd	**0.93 ±0.04**	**0.91 ±0.03**	**0.92 ±0.03**	**0.91 ±0.03**	**0.92 ±0.03**	**0.87 ±0.03**

Table 4.1: Effects of different regularisation methods, i.e *pre-trained weights*, *augmentation* (*aug*), and *weight decay* (*wd*), on the PH detection with the spatial PSAX-P model, when training on 10 random frames per subject. The best results are highlighted in **bold**.

	Percentile / # Frames	AUROC	F1-Score	Precision	Recall	Balanced Accuracy	Confidence
Max	95_{th}	0.87 ±0.05	0.85 ±0.03	0.87 ±0.03	0.85 ±0.03	0.85 ±0.04	0.89 ±0.02
	90_{th}	0.87 ±0.04	0.87 ±0.03	0.88 ±0.03	0.87 ±0.04	0.86 ±0.03	0.89 ±0.01
	80_{th}	0.90 ±0.06	0.88 ±0.05	0.89 ±0.04	0.88 ±0.05	0.88 ±0.05	0.86 ±0.02
	50_{th}	0.90 ±0.04	0.88 ±0.06	0.90 ±0.04	0.88 ±0.06	0.89 ±0.05	0.87 ±0.02
Min	95_{th}	0.93 ±0.05	0.90 ±0.06	0.91 ±0.04	0.89 ±0.06	0.90 ±0.05	0.89 ±0.02
	90_{th}	0.95 ±0.03	0.92 ±0.04	0.93 ±0.04	0.92 ±0.04	0.92 ±0.04	0.88 ±0.02
	80_{th}	**0.95 ±0.03**	**0.93 ±0.04**	**0.93 ±0.04**	**0.93 ±0.04**	**0.92 ±0.04**	**0.88 ±0.02**
	50_{th}	0.94 ±0.04	0.91 ±0.05	0.92 ±0.04	0.91 ±0.05	0.90 ±0.05	0.90 ±0.02
Rand	10	0.93 ±0.04	0.91 ±0.03	0.92 ±0.03	0.91 ±0.03	0.92 ±0.03	0.87 ±0.03
	all	0.94 ±0.03	0.91 ±0.03	0.92 ±0.03	0.91 ±0.03	0.90 ±0.04	0.86 ±0.02

Table 4.2: Results of PH detection with the spatial PSAX-P model, when varying the number of frames being trained on, and their extraction methods, i.e. maximum-expansion (*Max*), minimum-expansion (*Min*) or random (*Rand*). 95_{th}, 90_{th}, 80_{th} and 50_{th} percentile correspond on average to 7, 12, 26, and 60 frames per ECHO. The best results are highlighted in **bold**.

we get an F1-score of 0.90, almost the same as when selecting all frames (approx. 122 frames per ECHO). However, when selecting top 95_{th} maximum-expansion frames per video, the F1-score is only 0.85. This indicates that the minimum-expansion frames are the most informative frames. It is also consistent with the fact that PH-related abnormalities in the shape of the IVS and LV are most visible during systole, i.e. minimum-expansion of the heart [17]. Although reversed volume of the ventricles during diastole is also associated with PH, these results suggest that the systolic changes in LV and IVS morphology are more discriminative features for the model, as further verified by the interpretability analysis in Section 4.4.1.

The best results for the PSAX-P view are achieved when extracting around 26 minimum-expansion frames per ECHO (minimum 80_{th} percentile), resulting in an F1-score of 0.93. For future experiments, however, training is performed on 10 random frames instead. This is because extraction of minimum expansion frames is not possible for all views, and because training on 10 random frames aligns with previous work. The performance difference is also negligible, with balanced accuracy being the same, and other metrics being only slightly reduced. By further looking at the confusion matrices for these two approaches, in Tables 4.3 and 4.4 respectively, we see that the model trained on random frames only mis-classifies one additional PH patient as healthy, compared to the best model, although it also mis-classifies two additional healthy subjects as having PH.

		True		
		PH	*Not PH*	Total
Pred	*PH*	$TP = 52$	$FP = 15$	67
	Not PH	$FN = 4$	$TN = 104$	108
	Total	56	119	175

Table 4.3: Confusion matrix of the spatial model trained on 10 random PSAX-P frames per patient, for the task of PH detection.

		True		
		PH	*Not PH*	Total
Pred	*PH*	$TP = 53$	$FP = 13$	66
	Not PH	$FN = 3$	$TN = 106$	109
	Total	56	119	175

Table 4.4: Confusion matrix of the spatial model trained on top 26 minimum-expansion PSAX-P frames, for the task of PH detection.

4.2.2 Multiple Views for PH Detection and Severity Prediction

The final study of the spatial method involves applying the best approach, as determined by the ablation analysis from the previous section, to the other views and to the task of PH severity prediction. In Table 4.5 we report (a) the performance of the binary PH detection from each of the five views, as well as the results from the two multi-view approaches; majority voting (MV) and feature-level fusion (FLF). Specifically, for both multi-view approaches, we show the results of joining the three main views PSAX-P, PLAX and A4C (MV-3, FLF-3), as well as joining all five views (MV-All, FLF-All). We similarly report in Table 4.5(b) the performance of the severity PH prediction for each of the views, and all the multi-view approaches.

The best results for each task are achieved when using a multi-view approach. Further important findings from the table are discussed below.

(a)

View	AUROC	F1-Score	Precision	Recall	Balanced Accuracy	Confidence
A4C*	0.87±0.04	0.83±0.04	0.85±0.03	0.83±0.04	0.83±0.03	0.87±0.03
PLAX	0.92±0.05	0.88±0.04	0.89±0.04	0.88±0.04	0.88±0.04	0.89±0.02
PSAX-P	0.93±0.04	0.91±0.03	0.92 ±0.03	0.91 ±0.03	0.92 ±0.03	0.87 ±0.03
PSAX-S	0.83±0.03	0.81±0.03	0.83±0.02	0.81±0.03	0.81±0.03	0.86±0.04
PSAX-A	0.86±0.04	0.85±0.03	0.85±0.03	0.85±0.03	0.84±0.03	0.87±0.02
MV-3	[0.91 ±0.02]	0.88 ±0.02	0.88 ±0.02	0.88 ±0.02	0.88 ±0.02	0.88 ±0.01
MV-All	[0.92 ±0.02]	0.90 ±0.02	0.91 ±0.01	0.90 ±0.02	0.90 ±0.01	0.87 ±0.02
FLF-3	0.93 ±0.02	0.91 ±0.03	0.92 ±0.02	0.90 ±0.03	0.92 ±0.03	0.90 ±0.02
FLF-All	**0.95 ±0.04**	**0.93 ±0.03**	**0.94 ±0.03**	**0.93 ±0.04**	**0.93 ±0.03**	**0.91 ±0.02**

(b)

View	AUROC	F1-Score	Precision	Recall	Balanced Accuracy	Confidence
A4C	0.79±0.04	0.75±0.05	0.77±0.04	0.75±0.06	0.67±0.06	0.84±0.03
PLAX	0.84±0.04	0.76±0.04	0.78±0.05	0.77±0.04	0.70±0.05	0.85±0.03
PSAX-P	0.83±0.03	0.81±0.02	0.82±0.03	0.81±0.03	0.74±0.03	0.83±0.03
PSAX-S	0.74±0.07	0.68±0.06	0.70±0.07	0.70±0.08	0.62±0.06	0.83±0.03
PSAX-A	0.80±0.03	0.75±0.04	0.76±0.04	0.76±0.05	0.66±0.04	0.84±0.03
MV-3	[0.84 ±0.03]	0.81 ±0.03	0.83 ±0.04	0.82 ±0.03	0.74 ±0.06	0.85 ±0.02
MV-All	**[0.84±0.03]**	**[0.82 ±0.03]**	**0.83 ±0.04**	**0.83 ±0.03**	**0.73 ±0.04**	**0.85 ±0.01**
FLF-3	0.86 ±0.02	0.81 ±0.02	0.83 ±0.03	0.81 ±0.02	0.75 ±0.04	0.86 ±0.01
FLF-All	0.87 ±0.03	0.81 ±0.03	0.82 ±0.02	0.80 ±0.03	0.76 ±0.03	0.86 ±0.02

Table 4.5: Results from the spatial approach for (a) PH binary detection and (b) PH severity prediction. *MV-3* refers to majority voting of PSAX-P, PLAX and A4C views, and *MV-All* to majority voting of all views. *FLF-3* refers to feature-level fusion of PSAX-P, PLAX and A4C views, and *FLF-All* to feature-level fusion of all views. The best results for each task have been highlighted in **bold**.

Single-View Approach Among the single-view methods, the PSAX-P view shows the best performance for both tasks. Although the A4C view is one of the most commonly used views for cardiovascular disease diagnosis, our evaluation shows that it is not as discriminating as the PSAX-P or PLAX views. This is also in line with the neonatal echocardiography teaching manual [17], where it is stated that subjective assessment of PH from the A4C view in a 2D ECHO is usually only possible for moderate to severe PH cases, and quantitative evaluation is difficult. Furthermore, these results show that the PH severity prediction problem is more challenging than then PH detection, as the best single-view results for the PH detection yield an F1-score of 0.91, compared to only 0.81 for PH severity detection. This is to be expected, not only due to the harder task at hand but also because of the increased data imbalance. In this case, the robustness and accuracy can be increased by utilising more views.

Multi-View Approach Using multiple views does generally improve performance, and for both multi-view approaches combining all five views gives better performance than combining only the three most significant views.

For the task of PH severity prediction, all four multi-view approaches show some improvement in performance compared to when training only on a single view. The majority voting of all views (MV-All) seems to give the best results overall, although feature-level-fusion of all views (FLF-All) gives better balanced accuracy, and the AUROC of these two approaches is not directly comparable. The performance gain from using multiple views is still minimal, with for example the F1-score increasing only from 0.81 to 0.82. As we will see in Section 4.3, in the case of the spatio-temporal approach, majority voting provides a greater performance increase.

Using multiple views to enhance performance is less beneficial for the task of binary PH detection, as the performance is already competitive when using only a single view. And indeed the predictions for the PH detection do not improve when using the ensemble methods (MV-3, MV-ALL). When model performance is generally high, and one model of an ensemble performs significantly better than the others, like in our case, it is not uncommon that the ensemble performs no better than the best-performing model [58]. Combining all views with feature-level fusion (FLF-All) does however improve the results of the PH detection.

Comparison to State-of-the-Art The state-of-the-art approach for PH detection in adults [75], a spatial CNN trained on a single ECHO view (A4C), achieves an AUROC score of 0.85. With our spatial approach on the A4C view, we achieve similar results, or AUROC of 0.87, as seen in Table 4.5(a). Since both approaches involve training and evaluating a spatial CNN on 10 random frames from the A4C view, the similar results are expected and encouraging. It suggests that our data augmentation is effective, as our training dataset is considerably smaller than the dataset of [75]. However, we recognise that direct comparison is difficult, since the methods have been trained and evaluated on different datasets from a different population. Finally note that we improve on the state-of-the-art approach, by first of all training on the PSAX-P view, which significantly improves performance (AUROC 0.93), and second of all by incorporating temporal features, as seen in the next section 4.3.2, achieving an AUROC of 0.95 for the PH detection.

4.3 Spatio-Temporal Approach

In this section, we report the results of the spatio-temporal approach for the detection and severity prediction of PH. The insight gained from the ablation analysis of the spatial approach is applied to the spatio-temporal method, wherever applicable. This includes training on 10 samples (here sequences) per ECHO, and using all three regularisation techniques. Furthermore, in subsection 4.3.1 we perform additional ablation studies for determining specific spatio-temporal properties, by using the binary PH detection task and the base view (PSAX-P). In Subsection 4.3.2 we then provide an empirical assessment of the best spatio-temporal approach (as per the ablation analysis), for both the PH detection and the severity prediction, for all five views, and for the multi-view approach.

4.3.1 PH Detection on the PSAX-P View - Ablation

To understand the effects of different input sequence-lengths and different spatio-temporal architectures on performance, ablation studies were performed on the PSAX-P view, for the binary PH detection.

Sequence-length and Sampling Interval

We report in Table 4.6 the effects of varying the effective length (l) of the input sequences from min 8 to max 24, for sampling intervals s=1 and s=2, when training a ResNet3D-18 model on the PSAX-P view, for the task of binary PH detection. These settings correspond to sequence-lengths (k) in the range of 4 to 24. The best results are achieved when training on input sequences of length 12, where every consecutive frame is selected (i.e. l=12, k=12, s=1). Recall that a single heartbeat covers on average 10 frames, so when the input sequences cover 12 frames, they contain on average at least one heart-beat. The second-best performance is also achieved with sequences of length $k = 12$, but by sampling every other frame, such that effective length is 24 frames (i.e. l=24, k=12, s=2). Training on sequences with the same effective length yields in general rather similar performance. For sequences with effective length less than 16, sampling every frame (s=1) gives slightly improved performance compared to sampling every other frame (s=2). However, the opposite is true for sequences with effective length of 16 or longer. In that case, using a larger sampling interval and less number of frames (i.e. shorter sequence-length k), is beneficial. This is especially true for effective length of 24, where including each of the 24 frames in the sequence only gives an F1-score of 0.88, compared to 0.92 when sampling every other frame and thus only training on sequences of length 12. This might suggest that the model has difficulties in learning long-range dependencies.

k, s (l)	AUROC	F1-Score	Precision	Recall	Balanced Accuracy	Confidence
8, 1 (8)	0.92 ±0.03	0.90 ±0.04	0.92 ±0.02	0.90 ±0.04	0.91 ±0.03	0.90 ±0.02
12, 1 (12)	**0.95±0.04**	**0.92±0.03**	**0.93±0.03**	**0.92±0.03**	**0.94±0.03**	**0.90±0.03**
16, 1 (16)	0.92 ±0.04	0.90 ±0.04	0.91 ±0.03	0.89 ±0.04	0.91 ±0.04	0.92 ±0.02
24, 1 (24)	0.92 ±0.04	0.88 ±0.03	0.90 ±0.02	0.87 ±0.03	0.89 ±0.03	0.93 ±0.01
4, 2 (8)	0.92 ±0.04	0.89 ±0.05	0.91 ±0.03	0.89 ±0.05	0.90 ±0.04	0.89 ±0.02
6, 12 (12)	0.92 ±0.03	0.91 ±0.03	0.92 ±0.02	0.91 ±0.03	0.92 ±0.03	0.90 ±0.01
8, 2 (16)	0.92 ±0.04	0.91 ±0.03	0.92 ±0.03	0.91 ±0.03	0.91 ±0.03	0.92 ±0.02
12, 2 (24)	0.94 ±0.03	0.92 ±0.03	0.92 ±0.03	0.92 ±0.03	0.92 ±0.04	0.93 ±0.01

Table 4.6: Results of PH detection with the ResNet3D PSAX-P model when varying the effective length (l), sequence-length (k) and sampling interval (s) of the input sequences. The best results have been highlighted in **bold**.

Different 3D-CNN Architectures

In Table 4.7 we provide an empirical evaluation of the three different spatio-temporal architectures R(2+1)D, ResNet3D and SlowFast. In all cases, we train on 10 sequences per ECHO, with each sequence being of length 12 and with a sampling interval of 1.

Architecture	AUROC	F1-Score	Precision	Recall	Balanced Accuracy	Confidence
R(2+1)D, 18 layers	0.90 ±0.06	0.90 ±0.03	0.91 ±0.03	0.90 ±0.03	0.90 ±0.05	0.91 ±0.03
ResNet3D, 18 layers	0.95 ±0.04	0.92±0.03	0.93±0.02	0.92±0.03	0.94±0.03	0.90±0.03
SlowFast, 50 layers	0.93 ±0.04	0.90 ±0.04	0.91 ±0.04	0.90 ±0.04	0.90 ±0.05	0.91 ±0.01

Table 4.7: Performance of different spatio-temporal architectures when training with sequences of length 12 (k=12, s=1).

The 18-layer ResNet3D shows superior performance compared to the other two architectures. The other two yield similar results, with SlowFast being slightly better than R(2+1)D. Although the 50 layer SlowFast network has shown superior performances on various video classification tasks [18], it seems to lead to over-fitting on our relatively small dataset. Furthermore, factoring the 3D convolutional filters into separate spatial and temporal components, as in R(2+1)D, does not improve accuracy in our case.

4.3.2 Multiple Views for PH Detection and Severity Prediction

The final study involves applying the best spatio-temporal approach, as determined in the previous subsection, to all the views and for the task of PH severity prediction. That is, a ResNet3D-18 model, with input sequences consisting of 12 consecutive frames. For the multi-view approach, we only evaluate the ensemble method, i.e. majority voting, since training a spatio-temporal model with feature level fusion of five views is very memory heavy, with each input sample being of size 224 × 224 × 12 × 5. Additionally, such a large network would be very prone to overfitting on our relatively small dataset.

In Table 4.8 we report (a) the performance of the binary PH detection from each of the five views, as well as the results from majority voting of the three main views (MV-3) and all views (MV-All). We similarly report in Table 4.8(b) the performance of the severity PH prediction for each of the views, and the multi-view approaches. The best results for the PH severity prediction are achieved by majority voting of all views (same as for the spatial approach), but a single view model performs best on the PH detection task. Further important findings from the table are discussed below.

Single-View Approach Among the single-view methods, training on the PSAX-P view gives the best performance for both PH detection and severity prediction, followed by the PLAX view. The A4C view is not as discriminative as the other two main views. This is in line with the results from the spatial approach. Although not all views benefit from making use of the temporal domain, the two most significant views (PSASX-P and PLAX) do. For example, the PH severity F1-score increases from 0.76 to 0.78 when using a spatio-temporal model instead of a spatial model on the PLAX view. Similar to the spatial method, performance of the PH severity prediction is significantly lower than of the binary PH detection, but it can again be improved by utilising more views.

Multi-View Approach For the spatio-temporal approach, the performance of the PH severity prediction significantly improves when joining all views with majority voting (MV-All). The F1-score improves from 0.81 to 0.84 and balanced accuracy from 0.73 to 0.78. This gain is considerably larger than the gain achieved when applying majority voting to the spatial approach, which only improved F1-score by 0.1 and balanced accuracy was not improved. A possible reason is that for the spatio-temporal approach, the difference in performance between the two top-performing models (PSAX-P, PLAX) is very small (e.g. only 0.1 difference in balanced accuracy), but for the spatial approach, the difference is larger (0.4 difference in balanced accuracy). And as already mentioned, the greater the performance difference between the top-performing model and other models, the less likely an ensemble is to succeed.

Like for the spatial method, the ensemble does not improve results for the binary classification. Again, this is likely due to the fact that results are already quite competitive, and the performance difference between the top-performing model and other models is high. Here, we even see that the ensemble gives worse results than the top-performing model. When an ensemble performs worse than the best-performing member of the ensemble, it typically involves one top-performing model whose predictions are made worse by one or more poor-performing models and the ensemble is not able to effectively harness their contribution [58].

(a)

View	AUROC	F1-Score	Precision	Recall	Balanced Accuracy	Confidence
A4C	0.83±0.05	0.81±0.04	0.84±0.03	0.81±0.04	0.81±0.04	0.91±0.03
PLAX	0.90±0.07	0.86±0.09	0.88±0.07	0.86±0.09	0.86±0.08	0.91±0.02
PSAX-P	**0.95±0.04**	**0.92±0.03**	**0.93±0.03**	**0.92±0.03**	**0.94±0.03**	**0.90±0.03**
PSAX-S	0.79±0.04	0.81±0.03	0.82±0.04	0.81±0.03	0.80±0.04	0.90±0.02
PSAX-A	0.88±0.05	0.87±0.03	0.88±0.03	0.87±0.03	0.87±0.04	0.89±0.03
MV-3	[0.90±0.03]	0.87±0.04	0.88±0.03	0.87±0.04	0.87±0.04	0.92±0.01
MV-All	[0.90±0.03]	0.89±0.02	0.90±0.02	0.89±0.02	0.89±0.02	0.91±0.01

(b)

View	AUROC	F1-Score	Precision	Recall	Balanced Accuracy	Confidence
A4C	0.77±0.03	0.72±0.05	0.75±0.05	0.72±0.05	0.65±0.06	0.88±0.02
PLAX	0.85±0.04	0.78±0.05	0.82±0.06	0.79±0.06	0.72±0.05	0.89±0.03
PSAX-P	0.85±0.04	0.81±0.05	0.83±0.06	0.82±0.04	0.73±0.06	0.90±0.03
PSAX-S	0.73±0.07	0.68±0.08	0.69±0.09	0.69±0.08	0.62±0.07	0.85±0.04
PSAX-A	0.77±0.07	0.74±0.06	0.77±0.04	0.74±0.06	0.67±0.06	0.84±0.04
MV-3	[0.84±0.05]	0.83±0.05	0.86±0.04	0.83±0.05	0.76±0.07	0.91±0.02
MV-All	**[0.86±0.05]**	**0.84±0.06**	**0.86±0.05**	**0.85±0.05**	**0.78±0.07**	**0.89±0.02**

Table 4.8: Results from the spatio-temporal approach for (a) PH binary detection and (b) PH severity prediction. *MV-3* refers to majority voting of PSAX-P, PLAX and A4C views, and *MV-All* to majority voting of all views. The best results for each task are highlighted in **bold**.

4.4 Explainability

We provide a post-hoc explainability analysis of the single-view convolutions, showing the regions that are the most relevant for the assessment of PH. The analysis shows that the spatial and spatio-temporal models mostly highlight the same cardiac structures, and similarly the models for PH detection and severity prediction. We thus only report detailed results from the explainability analysis of the spatio-temporal models for PH severity prediction (in Subsection 4.4.1), to avoid duplication. However, as the salience maps do have slightly different characteristics for the spatial and spatio-temporal approaches, we briefly highlight those differences in Subsection 4.4.2. Further explainability results are found in Section A.3 in the Appendix, including saliency maps from models trained on the PH detection task.

4.4.1 Explainability of Spatio-temporal PH Severity Models

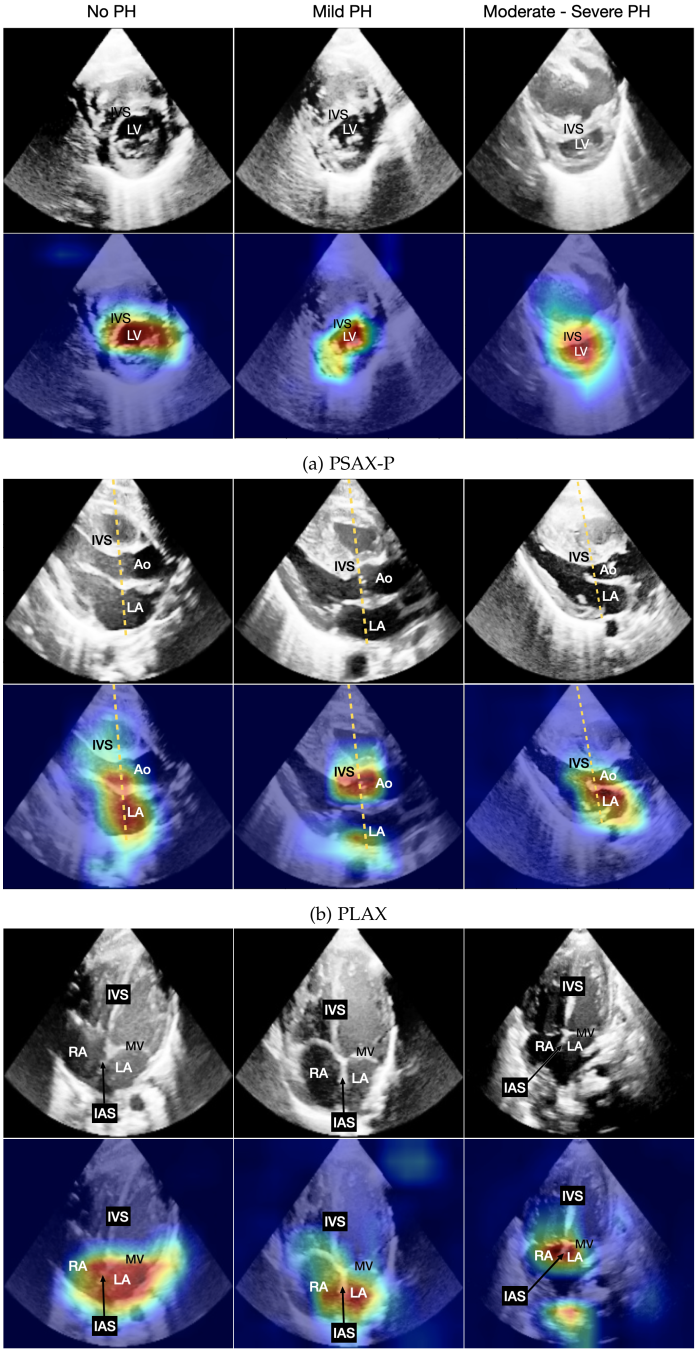
In Figure 4.1 we show an example of the spatio-temporal explainability analysis for three subjects with different levels of PH, for the three main views (a) PSAX-P, (b) PLAX and (c) A4C. In all cases, a single ECHO frame from each patient (top row) is combined with saliency maps using Grad-CAM (bottom row). Crucial cardiac structures highlighted by the saliency maps have been labelled with their abbreviation. These results show that our models are attending to clinically relevant cardiac structures, used for PH diagnosis by cardiologists. For each of the three main views, we provide further analysis of the attended features and relation to clinically relevant structures. Note that in a clinical setting, the visualisations can be viewed as a video containing spatio-temporal explanation. In the Appendix (Section A.3), we show an example of how the focus changes along the frames of a sequence.

PSAX-P

According to the neonatal echocardiography teaching manual [17], the ideal view for subjective evaluation of the interventricular septum (IVS) morphology and left ventricle (LV) shape is PSAX-P. In mild to moderate PH the IVS becomes flat during systole and in moderate to severe PH the septum bows into the LV, such that the LV becomes D-shaped, or crescentic. We show in Figure 4.1a that our PSAX-P severity prediction model focuses on the clinically relevant features that are recommended for diagnosis, that is the LV and IVS.

PLAX

Subjective evaluation of the IVS morphology is also possible from the PLAX view [17], and as we can see in Figure 4.1b the PLAX severity model does in-



(a) PSAX-P

(b) PLAX

(c) A4C

Figure 4.1: ECHO frames of subjects with no, mild and significant PH (top), combined with the saliency maps (bottom), for (a) PSAX-P view, (b) PLAX view and (c) A4C view. The yellow line shows how the LA:Ao is measured.

deed highlight the IVS area, among others. Furthermore, various quantitative assessments are frequently performed on the 2D ECHO of the PLAX view. This includes the left atrial-to-aortic root diameter ratio (LA:Ao), measured by extracting the M-mode, as demonstrated with the dashed yellow line in Figure 4.1b. We observe that the saliency maps highlight the pixels along the M-mode line, indicating that the PLAX model might be leveraging this ratio. Other common measurements of the PLAX view include measurements of the aortic valve (AV) annulus diameter, which is another area that the model focuses on. Overall, we see in Figure 4.1b that the model focuses on the areas around the LA, AV, Ao and IVS, suggesting that the model is able to consider both the relevant quantitative features and the subjective ones.

A4C

PH often leads to right ventricular (RV) dysfunction, which may cause RV dilatation and leftward deviation of the IVS. Thus, subjective evaluation of the A4C view typically involves examining the RV size and changes in IVS morphology. However, these changes are usually only visible for moderate to severe PH patients [17], and as we can see in Figure 4.1c the RV is not highlighted as an important feature for the A4C severity model. However, in patients with some degree of PH, parts of the IVS are highlighted. PH can furthermore lead to elevated right atrial pressures, which can be identified by a bulging interatrial septum (IAS) into the left atrium (LA) [17]. Although this is typically evaluated from the Atrical View, the features are also visible from the A4C view, and as seen in Figure 4.1c the A4C severity model mainly focuses on the LA and IAS area, suggesting these are the most discriminative features. The saliency maps also highlight the mitral valve (MV), and although MV regurgitation is predictive of PH, it is evaluated from a Doppler ECHO, and it is highly unlikely that it can be detected from the 2D ECHO.

4.4.2 Spatial vs. Spatio-temporal Saliency Maps

The analysis of the saliency maps for the binary PH detection models shows that the spatial models tend to attend to larger areas, compared to the spatio-temporal models, which tend to attend to more narrow and concentrated areas. In Figure 4.2 we see an example from a single PH patient, where both the spatial and spatio-temporal models attend to the left ventricle and IVS, but the area of focus is larger for the spatial model. This is not surprising, as the spatial models are less confident.

Furthermore, there is a slight difference in the attention of spatial models trained on random frames compared to minimum-expansion frames, as seen in Figure 4.3. In both cases, the attention is not as focused as for the temporal models, but models trained on random frames tend to attend to both left and right ventricles, whereas the attention is focused on the left ventricle and IVS only in the case of models trained on minimum expansion frames. A possible explanation is that when training on random frames, the ratio of the ventricles can be a better predictor for PH than the LV and IVS shape, as the shape deformation is mainly visible from minimum expansion frames.

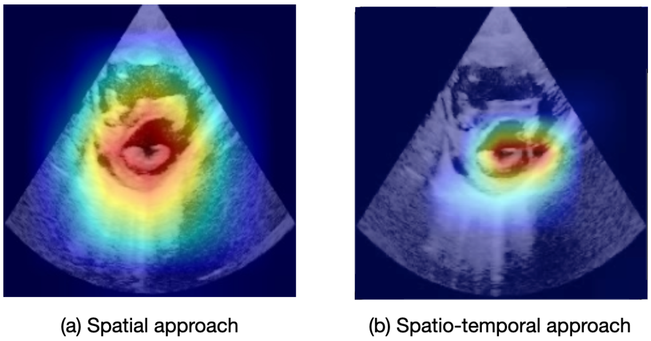

(a) Spatial approach
(b) Spatio-temporal approach

Figure 4.2: ECHO frame of a PH patient combined with the saliency maps from (a) spatial model trained on 10 random samples, and (b) spatio-temporal model trained on 10 random samples. Both models were trained for binary PH detection on the PSAX-P view.

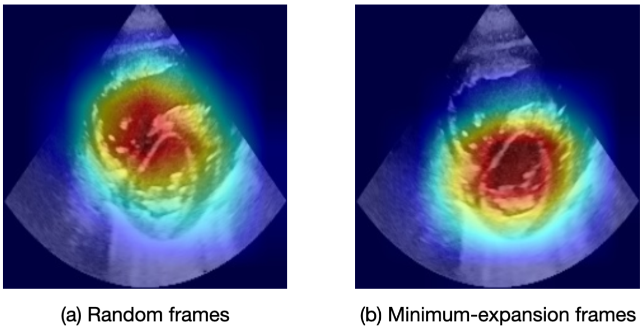

(a) Random frames
(b) Minimum-expansion frames

Figure 4.3: ECHO frame of a PH patient combined with the saliency maps from a spatial model trained for binary PH detection on the PSAX-P view, using (a) 10 random frames for training, and (b) top 80_{th} percentile minimum expansion frames

Chapter 5

Conclusions

In this work we developed an automated and streamlined method to assist clinicians in the assessment of PH in newborns, using either a spatial or spatio-temporal approach of one or more views. For the binary detection of PH we achieved optimal performance when using a spatio-temporal approach on the PSAX-P view or a spatial approach jointly trained on five views (i.e. the results of these two methods were comparable). For the severity prediction of PH we achieved optimal performance when performing majority voting of five spatio-temporal models, each trained on one of the five views. The advantage of the spatial approach is faster training time and less memory requirement, which also allows for relatively fast joint training of multiple views. The disadvantage is that for optimal performance, training should be performed on frames corresponding to minimum expansion of the heart, which requires segmentation of ventricles. For the spatio-temporal approach, no segmentation is required, and furthermore, the confidence of the spatio-temporal models was in general higher. This might be because all three dimensions of the ECHO are utilized to make a decision. Although the best single-view results are achieved with the spatio-temporal approach, some views did not benefit from adding the temporal domain. Additionally, a spatio-temporal model has a larger number of trainable parameters and is thus more prone to overfitting on small datasets. However, the fact that our spatio-temporal method performs comparable to or better than the spatial method on our relatively small dataset is encouraging, and we assume with more data, the benefits will be even larger.

The severity estimation of PH from ECHOs is critically important as it determines the urgency of treatment [10, 20], but it remains a challenge for cardiologist [12, 19]. Thus, our approach may have a considerable clinical impact in increasing the accuracy and steadiness of ECHO examinations by reducing the number of late or missed diagnoses of PH. Furthermore, it may assist less trained specialists and thereby reduce the workload of highly trained experts. Finally, by highlighting the input features that are crucial for the PH assessment, the proposed approach provides interpretable explanations for the clinicians, which in turn makes the system accountable.

A limitation of our method is the relatively small training dataset with only few severe PH patients, as well as the lack of an external test dataset. Furthermore, the annotations are based on visual inspection of a single cardiologists, but having the annotations verified by a second clinician would be beneficial. A dataset where PH has been determined invasively, with right-heart catheterization (RHC), would be the gold standard, but it is difficult to get around, as RHC is rarely used for initial PH assessment, but rather to confirm a diagnosis [47]. Future work includes increasing the clinical usability even further, by integrating our method into an end-to-end visualisation tool that can be used in a clinical setting. Before adoption of the tool for clinical routine, a performance evaluation using an external medical dataset is crucial, including a detailed analysis of failure cases. Finally, the addition and integration of additional information, such as the age or weight of the newborns, might be interesting.

Appendix A

Appendix

A.1 Implementation Details

The method was developed using the Python programming language, version 3.9.7, with the PyTorch deep learning library, version 1.10.1. Histogram equalization was performed with the OpenCV computer vision library, version 4.5.3. Metrics were calculated using the Scikit-learn machine learning library, version 0.24.2.

Experiments were run on a cluster containing different NVIDIA GeForce graphic cards: GTX 1080, GTX 1080 Ti, RTX 2080 Ti with 2048 MB RAM per processor core. Spatial models are trained with a single GPU, but spatio-temporal models with two GPUs. The multi-GPU training was implemented using the DataParlell module of PyTorch. Each model was trained for around 150 epochs per view minimizing the (categorical) cross-entropy loss with the Adam [32] optimiser. The single-view spatial models take around 3 hours to train with a batch size of 16, whereas the single-view spatio-temporal models take around 11 hours to train, with a batch size of 8. As a comparison, the multi-view spatio-temporal model of three views (i.e. feature-level fusion model) takes over 30 hours to train, and uses a batch size of 4. The smaller batch sizes are required for more complex models, due to memory constraints. Note that for the multi-view approach using majority voting, single-view models can be trained in parallel, and the training time is thus the same as for single-view models.

To ensure the reproducibility of this work, the code of IP-PHN was made publicly available under `https://anonymous.4open.science/r/echo_classification-DE4E/`

A.2 Dataset

A summary of the PH severity prediction dataset is provided below.

Feature	Value
PH (#None (%) / #Mild(%) / #Significant(%))	126(65%) / 32(16%) / 36(19%)
Age (days) (Mean ± SD)	56 ± 160
Maturity in birth (days) (Mean ± SD)	230 ± 46
Patient's weight (kg) (Mean ± SD)	2.9 ± 1.5
Manufacturer (Ultrasound Machine / Transducer)	GE Logic S8 / S4-10 at 6 MHz
Spatial size of original 2D images (pixels)	1440 x 866
Video length (frames)	122 ± 2
Video FPS	25 fps

Table A.1: Characteristics of the PH severity prediction dataset.

A.3 Explainability

In Figure A.1 we can see how the attention of the spatio-temporal models changes over time, for the task of PH severity prediction.

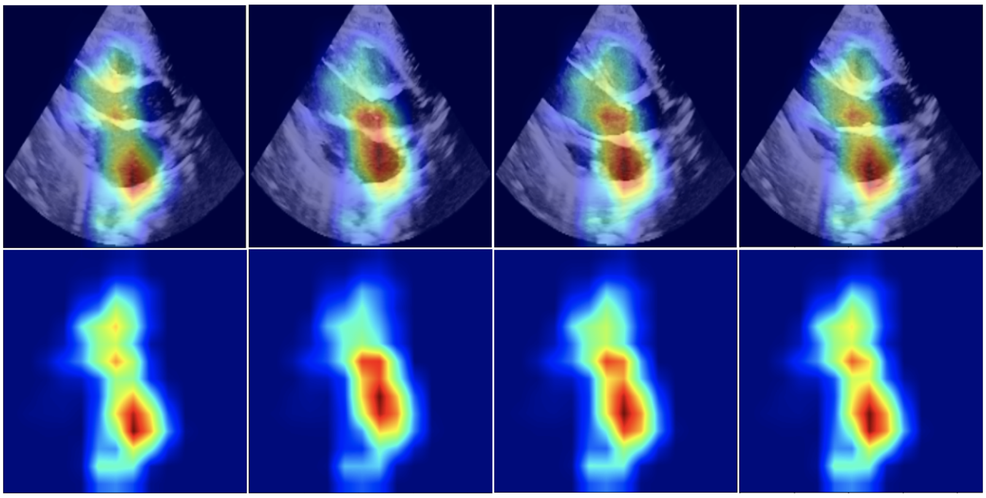


Figure A.1: Spatio-temporal Grad-CAM saliency maps (bottom) imposed on the original frames (top) for frames corresponding to systole, mid, diastole, mid in a PLAX ECHO.

A few examples of saliency maps for models trained on the task of PH binary detection are shown in Figure A.2 (for spatial models trained on random frames), Figure A.3 (for spatial models trained on minimum-expansion frames), and Figure A.4 (for spatio-temporal models)

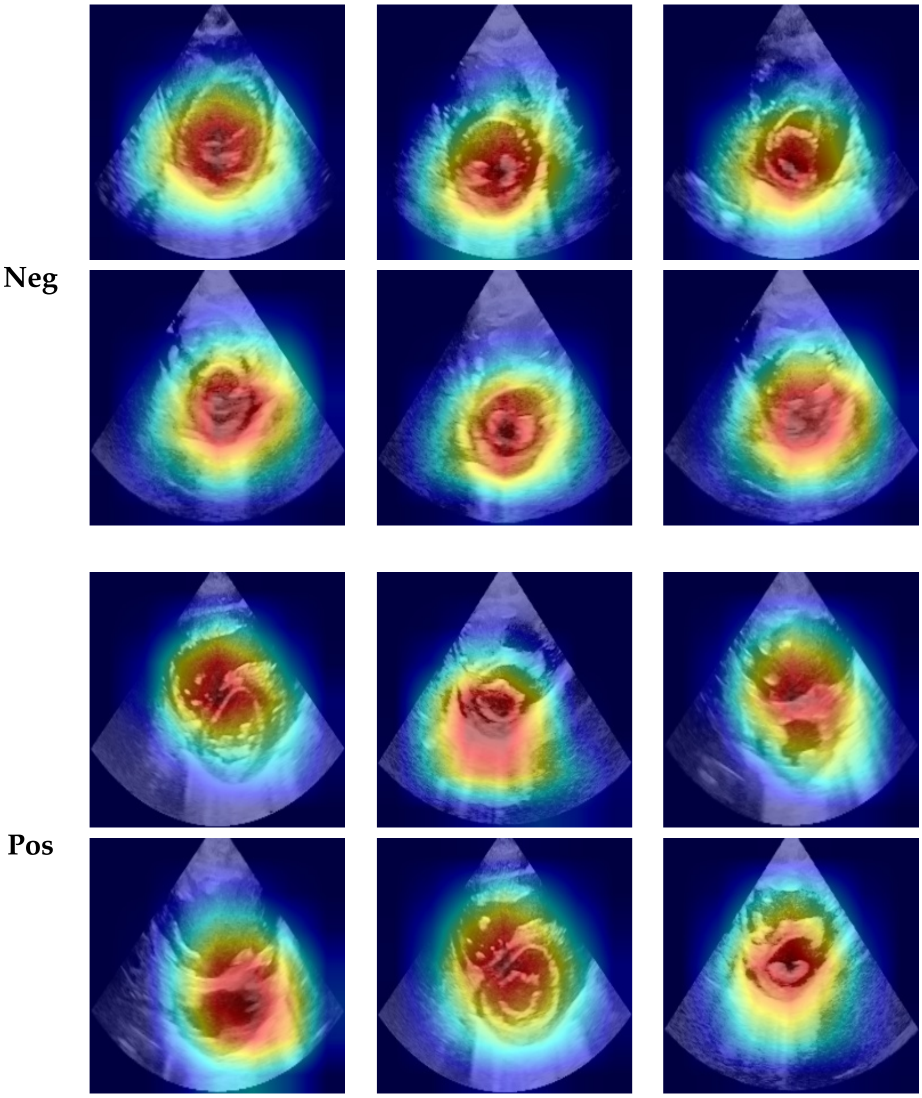


Figure A.2: Examples of Grad-CAM saliency maps imposed on original frames, for the spatial model trained on random frames of the PSAX-P view, for the task of binary PH prediction. First two rows are from healthy subjects with no PH (Neg), last two rows are from subjects with PH (Pos).

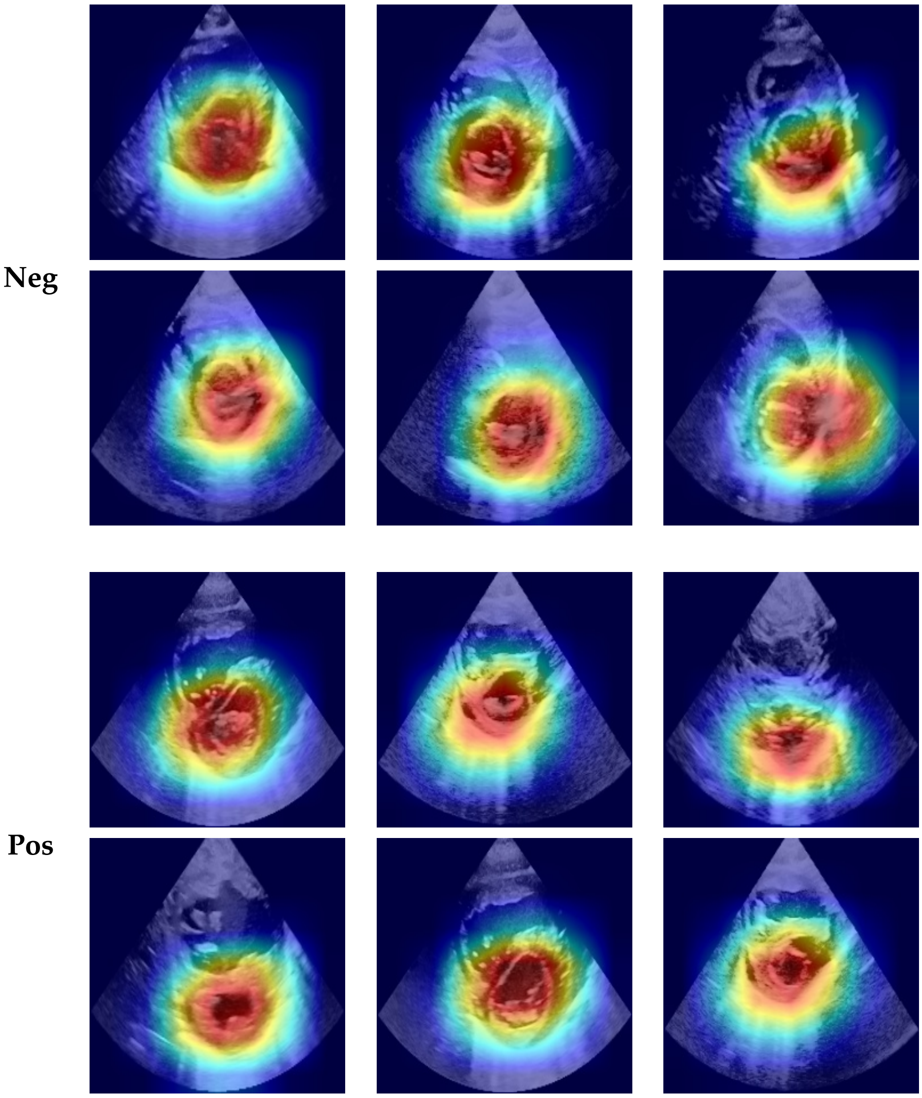


Figure A.3: Examples of Grad-CAM saliency maps imposed on original frames, for the spatial model trained on minimum-expansion frames of the PSAX-P view, for the task of binary PH prediction. First two rows (Neg) are from non-PH subjects, last two rows (Pos) are from subjects with PH.

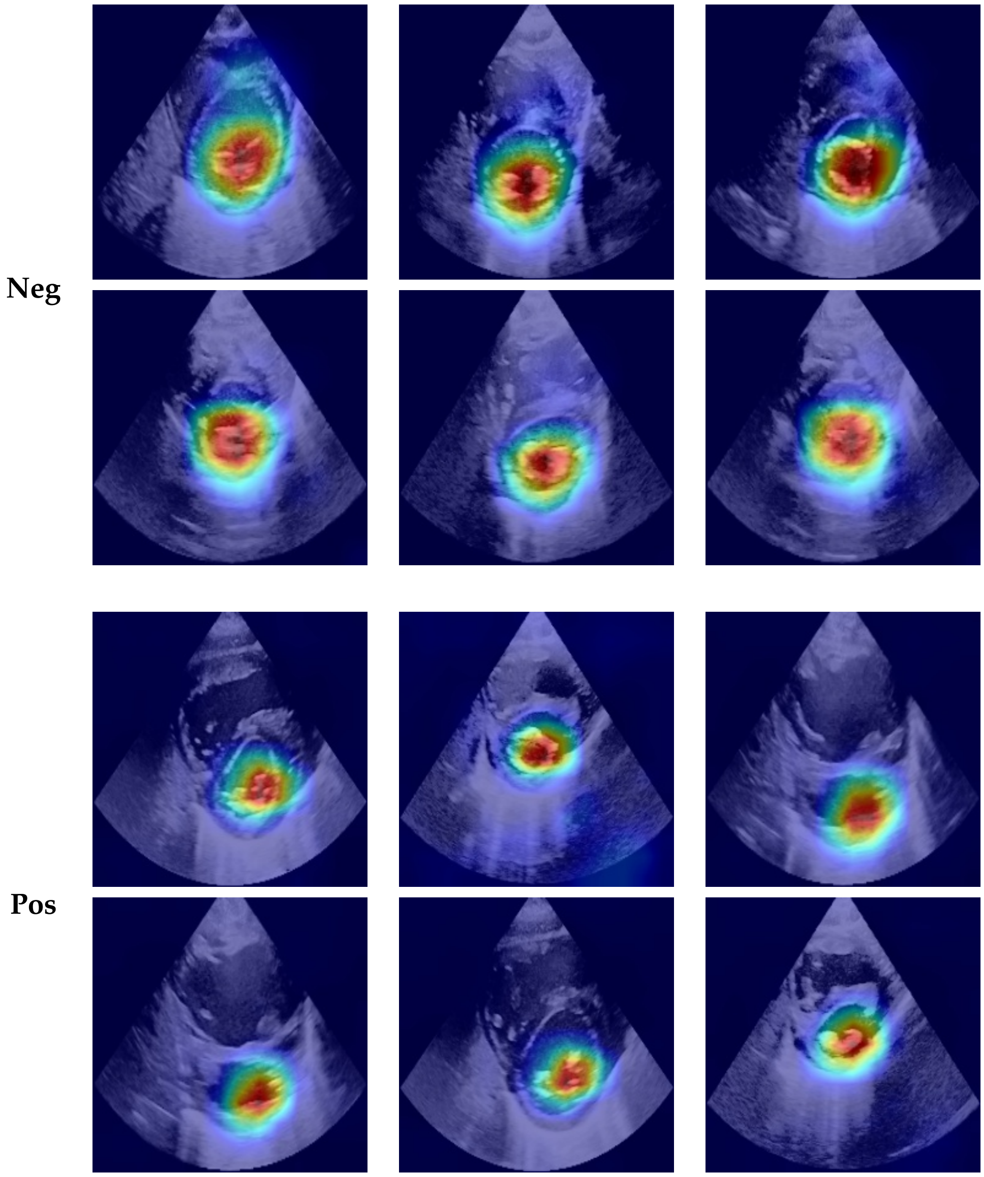


Figure A.4: Examples of Grad-CAM saliency maps imposed on original frames, of the spatio-temporal model trained on PSAX-P view, for the task of binary PH prediction. First two rows (Neg) are from non-PH subjects, last two rows (Pos) are from subjects with PH.

A.4 Additional Results

In this section we show sample-level performance of the main models for both PH detection and PH severity prediction, as well as the confusion matrices for both tasks.

A.4.1 Sample-level Performance

In Tables A.2 and A.3 we show the sample-level results for each of the views for the spatial approach and spatio-temporal approach, respectively. As expected, the sample-level scores are in general lower than the subject-level scores. The spatio-temporal approach achieves better results when looking at sample-level performance for all views, on the PH binary detection task. On the PH severity task it varies between views.

(a)

View	AUROC	F1-Score	Precision	Recall	Balanced Accuracy
A4C	0.82 ±0.04	0.76 ±0.03	0.77 ±0.03	0.76 ±0.04	0.74 ±0.04
PLAX	0.86 ±0.05	0.80 ±0.05	0.80 ±0.05	0.80 ±0.05	0.79 ±0.05
PSAX-P	0.88 ±0.04	0.82 ±0.04	0.83 ±0.04	0.82 ±0.04	0.81 ±0.05
PSAX-S	0.77 ±0.03	0.73 ±0.03	0.75 ±0.03	0.73 ±0.03	0.72 ±0.03
PSAX-A	0.82 ±0.04	0.77 ±0.03	0.78 ±0.03	0.77 ±0.03	0.75 ±0.03

(b)

View	AUROC	F1-Score	Precision	Recall	Balanced Accuracy
A4C	0.75 ±0.04	0.67 ±0.04	0.69 ±0.03	0.67 ±0.05	0.57 ±0.04
PLAX	0.80 ±0.04	0.70 ±0.03	0.71 ±0.03	0.71 ±0.04	0.60 ±0.04
PSAX-P	0.79 ±0.02	0.71 ±0.01	0.72 ±0.01	0.71 ±0.03	0.60 ±0.02
PSAX-S	0.71 ±0.06	0.62 ±0.04	0.63 ±0.05	0.63 ±0.06	0.53 ±0.05
PSAX-A	0.76 ±0.03	0.66 ±0.04	0.66 ±0.03	0.68 ±0.05	0.54 ±0.04

Table A.2: **Sample-wise** (i.e. per sequence) results from the spatial approach for (a) PH binary detection and (b) PH severity prediction.

A.4.2 Confusion Matrices

We report the confusion matrices of the main models trained on each of the two tasks, PH detection and PH severity prediction. Due to limited number of subjects per fold, we calculate the confusion matrices based on the concatenation of the 10 validation sets, resulting from the 10 folds. We make sure that each subject is only evaluated with the model that did not train on that subject. Note that this is different from other metrics, which are calculated per fold, and then the average over all the folds is reported.

The confusion matrices for the PH severity prediction task are shown in Tables A.4 and A.5, for the spatial and spatio-temporal models, respectively. The temporal model performs overall better than the spatial model, and it mis-classifies only one patient with significant PH as healthy. It however wrongly classifies 10 subjects with no or mild PH as having significant PH, which is more than the spatial model. The temporal model rather struggles with distinguishing between no PH and mild PH.

The confusion matrices for the binary PH detection task are shown in Tables A.6 and A.8, for the spatial and temporal model, respectively. Additionally, in Table A.7, we show the confusion matrix of the spatial model trained on minimum-expansion frames instead of random frames. The temporal model performs significantly best, only misclassifying one PH-patient as healthy (but wrongly classifying 10 healthy subjects as having PH).

(a)

View	AUROC	F1-Score	Precision	Recall	Balanced Accuracy
A4C	0.80 ±0.05	0.77 ±0.04	0.78 ±0.04	0.76 ±0.04	0.75 ±0.04
PLAX	0.88 ±0.07	0.81 ±0.07	0.83 ±0.06	0.81 ±0.07	0.80 ±0.06
PSAX-P	0.91 ±0.04	0.86 ±0.03	0.86 ±0.03	0.86 ±0.03	0.85 ±0.03
PSAX-S	0.77 ±0.03	0.75 ±0.03	0.75 ±0.03	0.75 ±0.03	0.72 ±0.03
PSAX-A	0.84 ±0.04	0.80 ±0.03	0.81 ±0.03	0.80 ±0.03	0.79 ±0.04

(b)

View	AUROC	F1-Score	Precision	Recall	Balanced Accuracy
A4C	0.75 ±0.02	0.67 ±0.05	0.69 ±0.05	0.67 ±0.05	0.58 ±0.06
PLAX	0.82 ±0.04	0.72 ±0.05	0.75 ±0.05	0.73 ±0.06	0.65 ±0.06
PSAX-P	0.84 ±0.04	0.75 ±0.04	0.76 ±0.04	0.76 ±0.04	0.65 ±0.05
PSAX-S	0.70 ±0.07	0.63 ±0.07	0.63 ±0.07	0.63 ±0.07	0.54 ±0.06
PSAX-A	0.75 ±0.06	0.66 ±0.04	0.68 ±0.04	0.65 ±0.04	0.56 ±0.04

Table A.3: **Sample-wise** (i.e. per sequence) results from the spatio-temporal approach for (a) PH binary detection and (b) PH severity prediction.

		True			
		No PH	*Mild PH*	*Sign. PH*	Total
Pred	*No PH*	104	12	2	117
	Mild PH	10	14	3	31
	Sign. PH	2	3	23	28
	Total	117	31	28	176

Table A.4: Confusion matrix of the PSAX-P spatial model trained on random frames, when training on the task of PH severity prediction.

		True			
		No PH	*Mild PH*	*Sign. PH*	Total
Pred	*No PH*	110	7	1	118
	Mild PH	3	18	3	24
	Sign. PH	4	6	24	34
	Total	117	31	28	176

Table A.5: Confusion matrix of the PSAX-P spatio-temporal model, when training on the task of PH severity prediction.

		True		
		Positive	*Negative*	Total
Pred	*Positive*	$TP = 52$	$FP = 15$	67
	Negative	$FN = 4$	$TN = 104$	108
	Total	56	119	175

Table A.6: Confusion matrix of the PSAX-P spatial model trained on 10 random frames, for the task of PH detection.

		True		
		Positive	*Negative*	Total
Pred	*Positive*	$TP = 53$	$FP = 13$	66
	Negative	$FN = 3$	$TN = 106$	109
	Total	56	119	175

Table A.7: Confusion matrix of the PSAX-P spatial model trained on minimum-expansion frames, for the task of PH detection.

		True		
		Positive	*Negative*	Total
Pred	*Positive*	$TP = 55$	$FP = 10$	65
	Negative	$FN = 1$	$TN = 109$	110
	Total	56	119	175

Table A.8: Confusion matrix of the PSAX-P spatio-temporal model, for the task of PH detection.

Bibliography

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9525–9536, Red Hook, NY, USA, 2018. Curran Associates Inc.

[2] J. Arunnehru, G. Chamundeeswari, and S. Bharathi. Human action recognition using 3d convolutional neural networks with 3d motion cuboids in surveillance videos. *Procedia Computer Science*, 133:471–477, 01 2018.

[3] Daniel X Augustine, Lindsay D Coates-Bradshaw, James Willis, Allan Harkness, Liam Ring, Julia Grapsa, Gerry Coghlan, Nikki Kaye, David Oxborough, Shaun Robinson, Julie Sandoval, Bushra S Rana, Anjana Siva, Petros Nihoyannopoulos, Luke S Howard, Kevin Fox, Sanjeev Bhattacharyya, Vishal Sharma, Richard P Steeds, Thomas Mathew, and _ _. Echocardiographic assessment of pulmonary hypertension: a guideline protocol from the british society of echocardiography. *Echo Research and Practice*, 5(3):G11–G24, September 2018.

[4] Robyn J. Barst, Michael D. McGoon, C. Gregory Elliott, Aimee J. Foreman, Dave P. Miller, and D. Dunbar Ivy. Survival in childhood pulmonary arterial hypertension. *Circulation*, 125(1):113–122, January 2012.

[5] Ghalib A. Bello, Timothy J. W. Dawes, Jinming Duan, Carlo Biffi, Antonio de Marvao, Luke S. G. E. Howard, J. Simon R. Gibbs, Martin R. Wilkins, Stuart A. Cook, Daniel Rueckert, and Declan P. O'Regan. Deep-learning cardiac motion analysis for human survival prediction. *Nature Machine Intelligence*, 1(2):95–104, February 2019.

[6] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[7] Nicolas Brugger, Mona Lichtblau, Micha Maeder, Hajo Muller, Cyril Pellaton, Patrick Yerly, and Swiss Society for Pulmonary Hypertension SSPH. Two-dimensional transthoracic echocardiography at rest for the diagnosis, screening and management of pulmonary hypertension. *Swiss Medical Weekly*, June 2021.

[8] Christoph B. Burckhardt. Speckle in ultrasound b-mode scans. *IEEE Transactions on Sonics and Ultrasonics*, 25(1):1–6, 1978.

[9] J. Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. pages 4724–4733, 07 2017.

[10] Paul Corris and Bruno Degano. Severe pulmonary arterial hypertension: treatment options and the bridge to transplantation. *European Respiratory Review*, 23(134):488–497, December 2014.

[11] Michele D'Alto, Emanuele Romeo, Paola Argiento, Antonello D'Andrea, Rebecca Vanderpool, Anna Correra, Eduardo Bossone, Berardo Sarubbi, Raffaele Calabrò, Maria Giovanna Russo, and Robert Naeije. Accuracy and precision of echocardiography versus right heart catheterization for the assessment of pulmonary hypertension. *International Journal of Cardiology*, 168(4):4058–4062, October 2013.

[12] Soham Dasgupta, Joan C. Richardson, Ashraf M. Aly, and Sunil K. Jain. Role of functional echocardiographic parameters in the diagnosis of bronchopulmonary dysplasia-associated pulmonary hypertension. *Journal of Perinatology*, 42(1):19–30, March 2021.

[13] Timothy J. W. Dawes, Antonio de Marvao, Wenzhe Shi, Tristan Fletcher, Geoffrey M. J. Watson, John Wharton, Christopher J. Rhodes, Luke S. G. E. Howard, J. Simon R. Gibbs, Daniel Rueckert, Stuart A. Cook, Martin R. Wilkins, and Declan P. O'Regan. Machine learning of three-dimensional right ventricular motion enables outcome prediction in pulmonary hypertension: A cardiac MR imaging study. *Radiology*, 283(2):381–390, May 2017.

[14] Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15. Springer Berlin Heidelberg, 2000.

[15] Jeff Donahue, Lisa Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. Long-term recurrent convolutional networks for visual recognition and description. pages 2625–2634, 06 2015.

[16] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell.

Long-term recurrent convolutional networks for visual recognition and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):677–691, 2017.

[17] Afif EL-Khuffash. Neonatal echocardiography teaching manual. 01 2014.

[18] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[19] Micah R. Fisher, Paul R. Forfia, Elzbieta Chamera, Traci Housten-Harris, Hunter C. Champion, Reda E. Girgis, Mary C. Corretti, and Paul M. Hassoun. Accuracy of doppler echocardiography in the hemodynamic assessment of pulmonary hypertension. *American Journal of Respiratory and Critical Care Medicine*, 179(7):615–621, April 2009.

[20] Nazzareno Galiè, Marc Humbert, Jean-Luc Vachiery, Simon Gibbs, Irene Lang, Adam Torbicki, Gérald Simonneau, Andrew Peacock, Anton Vonk Noordegraaf, Maurice Beghetti, Ardeschir Ghofrani, Miguel Angel Gomez Sanchez, Georg Hansmann, Walter Klepetko, Patrizio Lancellotti, Marco Matucci, Theresa McDonagh, Luc A. Pierard, Pedro T. Trindade, Maurizio Zompatori, and Marius Hoeper. 2015 ESC/ERS guidelines for the diagnosis and treatment of pulmonary hypertension. *European Respiratory Journal*, 46(4):903–975, August 2015.

[21] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:3681–3688, July 2019.

[22] Amirata Ghorbani, David Ouyang, Abubakar Abid, Bryan He, Jonathan H. Chen, Robert A. Harrington, David H. Liang, Euan A. Ashley, and James Y. Zou. Deep learning interpretation of echocardiograms. *npj Digital Medicine*, 3(1), January 2020.

[23] Sebastian Greiner, Andreas Jud, Matthias Aurich, Alexander Hess, Thomas Hilbel, Stefan Hardt, Hugo A. Katus, and Derliz Mereles. Reliability of noninvasive assessment of systolic pulmonary artery pressure by doppler echocardiography compared to right heart catheterization: Analysis in a large patient population. *Journal of the American Heart Association*, 3(4), August 2014.

[24] Georg Hansmann. Pulmonary hypertension in infants, children, and young adults. *Journal of the American College of Cardiology*, 69(20):2551–2569, 2017.

[25] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. pages 3154–3160, 10 2017.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[27] David J. Hur and Lissa Sugeng. Non-invasive multimodality cardiovascular imaging of the right heart and pulmonary circulation in pulmonary hypertension. *Frontiers in Cardiovascular Medicine*, 6, March 2019.

[28] Tarek Kaddoura, Karunakar Vadlamudi, Shine Kumar, Prashant Bobhate, Long Guo, Shreepal Jain, Mohamed Elgendi, James Y Coe, Daniel Kim, Dylan Taylor, Wayne Tymchak, Dale Schuurmans, Roger J. Zemp, and Ian Adatia. Acoustic diagnosis of pulmonary hypertension: automated speech- recognition-inspired classification algorithm outperforms physicians. *Scientific Reports*, 6(1), September 2016.

[29] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017.

[30] Dongyi Kim, Hyeon Cho, Hochul Shin, Soo-Chul Lim, and Wonjun Hwang. An efficient three-dimensional convolutional neural network for inferring physical interaction force from video. *Sensors*, 19(16):3579, August 2019.

[31] Gary King and Langche Zeng. Logistic regression in rare events data. *Political Analysis*, 9:137 – 163, 2001.

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[33] Kenya Kusunose, Yukina Hirata, Takumasa Tsuji, Jun'ichi Kotoku, and Masataka Sata. Deep learning to predict elevated pulmonary artery pressure in patients with suspected pulmonary hypertension using standard chest X-ray. *Scientific Reports*, 10(1), November 2020.

[34] Joon-Myoung Kwon, Kyung-Hee Kim, Jose Medina-Inojosa, Ki-Hyun Jeon, Jinsik Park, and Byung-Hee Oh. Artificial intelligence for early prediction of pulmonary hypertension using electrocardiography. *The Journal of Heart and Lung Transplantation*, 39(8):805–814, August 2020.

[35] Ricardo Bigolin Lanfredi, Ambuj Arora, Trafton Drew, Joyce D. Schroeder, and Tolga Tasdizen. Comparing radiologists' gaze and saliency maps generated by interpretability methods for chest x-rays, 2021.

[36] Hyebin Lee, Seong Tae Kim, and Yong Man Ro. Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, pages 21–29. Springer International Publishing, 2019.

[37] Andreas Leha, Kristian Hellenkamp, Bernhard Unsöld, Sitali Mushemi-Blake, Ajay M. Shah, Gerd Hasenfuß, and Tim Seidler. A machine learning approach for the prediction of pulmonary hypertension. *PLOS ONE*, 14(10):e0224453, October 2019.

[38] Dan Li, Kaifeng Zhang, Zhenbo Li, and Yifei Chen. A spatiotemporal convolutional network for multi-behavior recognition of pigs. *Sensors*, 20(8):2381, April 2020.

[39] Zhenqiang Li, Weimin Wang, Zuoyue Li, Yifei Huang, and Yoichi Sato. Towards visually explaining video understanding networks with perturbation. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, January 2021.

[40] Charles X. Ling and Chenghui Li. Data mining for direct marketing: Problems and solutions. In *KDD*, 1998.

[41] Kuan Liu, Yanen Li, N. Xu, and P. Natarajan. Learn to combine modalities in multimodal deep learning. *ArXiv*, abs/1805.11730, 2018.

[42] Carol Mitchell, Peter Rahko, Lori Blauwet, Barry Canaday, Joshua Finstuen, Michael Foster, Kenneth Horton, Kofo Ogunyankin, Richard Palma, and Eric Velazquez. Guidelines for performing a comprehensive transthoracic echocardiographic examination in adults: Recommendations from the american society of echocardiography. *Journal of the American Society of Echocardiography*, 32, 10 2018.

[43] Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022.

[44] Hiroki Mori, Kei Inai, Hisashi Sugiyama, and Yoshihiro Muragaki. Diagnosing atrial septal defect from electrocardiogram with deep learning. *Pediatric Cardiology*, 42(6):1379–1387, April 2021.

[45] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2016.

[46] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 689–696, Madison, WI, USA, 2011. Omnipress.

[47] Jin-Rong Ni, Pei-Jing Yan, Shi-Dong Liu, Yuan Hu, Ke-Hu Yang, Bing Song, and Jun-Qiang Lei. Diagnostic accuracy of transthoracic echocardiography for pulmonary hypertension: a systematic review and meta-analysis. *BMJ Open*, 9(12):e033084, December 2019.

[48] Jared M. O'Leary, Tufik R. Assad, Meng Xu, Eric Farber-Eger, Quinn S. Wells, Anna R. Hemnes, and Evan L. Brittain. Lack of a tricuspid regurgitation doppler signal and pulmonary hypertension by invasive measurement. *Journal of the American Heart Association*, 7(13), July 2018.

[49] David Opitz and Richard Maclin. Popular ensemble methods: An empirical study. 11(1):169–198, jul 1999.

[50] David Ouyang, Bryan He, Amirata Ghorbani, Matthew P. Lungren, Euan A. Ashley, David H. Liang, and James Y. Zou. Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning. 2019.

[51] David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph E. Ebinger, C. Langlotz, Paul A. Heidenreich, Robert A. Harrington, David H. Liang, Euan A. Ashley, and James Y. Zou. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580:252–256, 2020.

[52] Itthisak Phueaksri and Sukree Sinthupinyp. Convolutional neural network using stacked frames for video classification. In *Proceedings of the 2019 2nd International Conference on Computational Intelligence and Intelligent Systems*, CIIS 2019, page 85–89, New York, NY, USA, 2019. Association for Computing Machinery.

[53] Esther Puyol-Antón, Baldeep S. Sidhu, Justin Gould, Bradley Porter, Mark K. Elliott, Vishal Mehta, Christopher A. Rinaldi, and Andrew P. King. A multimodal deep learning model for cardiac resynchronisation therapy response prediction, 2021.

[54] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019.

[55] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S.

Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014.

[56] Matthias Schneider, Anna Maria Pistritto, Christian Gerges, Mario Gerges, Christina Binder, Irene M. Lang, Gerald Maurer, Thomas Binder, and Georg Goliasch. Multi-view approach for the diagnosis of pulmonary hypertension using transthoracic echocardiography. *The International Journal of Cardiovascular Imaging*, 34:695 – 700, 2017.

[57] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[58] Giovanni Seni and John Elder. *Ensemble methods in data mining*. Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan & Claypool, San Rafael, CA, February 2010.

[59] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 802–810, 2015.

[60] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2014.

[61] Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia - MULTIMEDIA '05*. ACM Press, 2005.

[62] Xiaolin Song, Cuiling Lan, Wenjun Zeng, Junliang Xing, Xiaoyan Sun, and Jingyu Yang. Temporal–spatial mapping for action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(3):748–759, March 2020.

[63] Jessica Torres Soto, J. Weston Hughes, Pablo Amador Sanchez, Marco Perez, David Ouyang, and Euan Ashley. Multimodal deep learning enhances diagnostic precision in left ventricular hypertrophy. June 2021.

[64] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. *CoRR*, abs/1412.6806, 2015.

[65] Robin H. Steinhorn. Neonatal pulmonary hypertension. *Pediatric Critical Care Medicine*, 11:S79–S84, 2010.

[66] Alexandros Stergiou, Georgios Kapidis, Grigorios Kalliatakis, Christos Chrysoulas, Ronald Poppe, and Remco Veltkamp. Class feature pyramids for video explanation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4255–4264, 2019.

[67] Alexandros Stergiou, Georgios Kapidis, Grigorios Kalliatakis, Christos Chrysoulas, Remco Veltkamp, and Ronald Poppe. Saliency tubes: Visual explanations for spatio-temporal convolutions. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1830–1834, 2019.

[68] Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery*, 10(5), June 2020.

[69] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, December 2015.

[70] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. pages 6450–6459, 06 2018.

[71] Tuomas Vainio, Teemu Mäkelä, Sauli Savolainen, and Marko Kangasniemi. Performance of a 3d convolutional neural network in the detection of hypoperfusion at CT pulmonary angiography in patients with chronic pulmonary embolism: a feasibility study. *European Radiology Experimental*, 5(1), September 2021.

[72] Lei Wang, Yangyang Xu, Jun Cheng, Haiying Xia, Jianqin Yin, and Jiaji Wu. Human action recognition by learning spatio-temporal features with deep neural networks. *IEEE Access*, 6:17913–17922, 2018.

[73] Saining Xie, Chen Sun, Jonathan Huang, Z. Tu, and Kevin Murphy. *Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*, pages 318–335. 09 2018.

[74] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pages 818–833. Springer International Publishing, 2014.

[75] Jeffrey Zhang, Sravani Gajjala, Pulkit Agrawal, Geoffrey Tison, Laura Hallock, Lauren Beussink, Mats Lassen, Eugene Fan, Mandar Aras, ChaRandle Jordan, Kirsten Fleischmann, Michelle Melisko, Atif Qasim, Sanjiv Shah, Ruzena Bajcsy, and Rahul Deo. Fully automated echocardiogram interpretation in clinical practice: Feasibility and diagnostic accuracy. *Circulation*, 138:1623–1635, 10 2018.

[76] Jeffrey Zhang, Sravani Gajjala, Pulkit Agrawal, Geoffrey H. Tison, Laura A. Hallock, Lauren Beussink-Nelson, Mats Christian Højbjerg Lassen, Eugene Fan, Mandar A. Aras, ChaRandle Jordan, Kirsten E. Fleischmann, Michelle E. Melisko, Atif Qasim, Alexei A. Efros, Sanjiv J. Shah, Ruzena Bajcsy, and Rahul C. Deo. A computer vision pipeline for automated determination of cardiac structure and function and detection of disease by two-dimensional echocardiography. *arXiv: Computer Vision and Pattern Recognition*, 2017.

[77] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.

[78] Guangming Zhu, Liang Zhang, Peiyi Shen, and Juan Song. Multimodal gesture recognition using 3-d convolution and convolutional LSTM. *IEEE Access*, 5:4517–4524, 2017.

[79] Peifei Zhu and Masahiro Ogino. Guideline-based additive explanation for computer-aided diagnosis of lung nodules. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, pages 39–47. Springer International Publishing, 2019.

[80] Xiao-Ling Zou, Yong Ren, Ding-Yun Feng, Xu-Qi He, Yue-Fei Guo, Hai-Ling Yang, Xian Li, Jia Fang, Quan Li, Jun-Jie Ye, Lan-Qing Han, and Tian-Tuo Zhang. A promising approach for screening pulmonary hypertension based on frontal chest radiographs using deep learning: A retrospective study. *PLOS ONE*, 15(7):e0236378, July 2020.

[81] Małgorzata Żuk, Katarzyna Mazurkiewicz-Antoń, Anna Migdał, Dorota Jagiełłowicz-Kowalska, Anna Turska-Kmieć, Lidia Ziółkowska, Grażyna Brzezińska-Rajszys, Maria Zubrzycka, and Wanda Kawalec. Prognosis in children with pulmonary arterial hypertension: 10-year single-centre experience. *Kardiologia Polska (Polish Heart Journal)*, 74(2):159 – 167, 2016.

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Interpretable Prediction of Pulmonary Hypertension in Newborns
using Echocardiography

Authored by (in block letters):
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
Ragnarsdóttir	Hanna

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
Uster, 21.03.2022	Hanna R

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.