applied sciences

MDPI

Article

Acute Respiratory Distress Identification via Multi-Modality Using Deep Learning

Wajahat Nawaz [1], Kevin Albert [2], Philippe Jouvet [2,*] and Rita Noumeir [1]

[1] Biomedical Information Processing Laboratory, École de Technologie Supérieure, University of Québec, Montreal, QC H3C 1K3, Canada
[2] Research Center at CHU Sainte-Justine Hospital, University of Montreal, Montreal, QC H3T 1J4, Canada
* Correspondence: philippe.jouvet.med@ssss.gouv.qc.ca

Abstract: Medical instruments are essential in pediatric intensive care units (PICUs) for measuring respiratory parameters to prevent health complications. However, the assessment of acute respiratory distress (ARD) is still conducted through intermittent visual examination. This process is subjective, labor-intensive, and prone to human error, making it unsuitable for continuous monitoring and early detection of deterioration. Previous studies have proposed solutions to address these challenges, but their techniques rely on color information, the performance of which can be influenced by variations in skin tone and lighting conditions. We propose leveraging multi-modality data to address these limitations. Our method integrates color and depth data using deep convolutional neural networks with a late feature fusion scheme. We train and evaluate our model on a dataset of 153 patients with respiratory illnesses, 86 of whom have ARD of varying severity levels. Experimental results demonstrate that multi-modality data combined with simple late fusion techniques are more effective with limited data, offering higher confidence scores compared to using color information alone. Our approach achieves an accuracy of 85.2%, a precision of 86.7%, a recall of 85.2%, and an F_1 score of 85.8%. These findings suggest that multi-modality data provide a promising solution for improving ARD detection accuracy and confidence in clinical settings.

Keywords: acute respiratory distress; action recognition; deep learning; multi-modality; transfer learning; two-stream network; video classification

Academic Editors: Stefano Nobile and Alessandro Perri

Received: 27 November 2024
Revised: 20 January 2025
Accepted: 28 January 2025
Published: 2 February 2025

Citation: Nawaz, W.; Albert, K.; Jouvet, P.; Noumeir, R. Acute Respiratory Distress Identification via Multi-Modality Using Deep Learning. *Appl. Sci.* **2025**, *15*, 1512. https://doi.org/10.3390/app15031512

1. Introduction

Acute respiratory distress (ARD) is a leading cause of infant admissions to the pediatric intensive care unit (PICU) [1]. This life-threatening condition is characterized by insufficient oxygen saturation levels in the bloodstream, often resulting from underlying lung diseases [2]. In response to ARD, the brain activates accessory respiratory muscles to ensure an adequate oxygen supply and maintain oxygen saturation in the bloodstream. However, prolonged overuse of these muscles can lead to fatigue and, ultimately, respiratory failure. Therefore, early detection of ARD is crucial for timely interventions, such as providing external respiratory support, to prevent severe health complications [1].

Patients with ARD exhibit several visible signs, including an elevated respiratory rate (RR), reduced oxygen saturation levels, a distressed appearance, thoracic-abdominal asynchrony (TAA), and chest retraction signs [3]. Traditionally, healthcare professionals evaluate these parameters through visual examinations. This process, which involves manually counting respiratory rate (RR) and observing signs like TAA, is labor-intensive, subjective, and prone to human error. While advancements in medical technology have

introduced devices such as respiratory inductance plethysmography (RIP) and pulse oximeters for real-time measurement of RR, TAA, and oxygen saturation levels, these methods are often uncomfortable for patients, requiring cooperation that can be challenging in children. Additionally, they can cause skin irritation, restrict movement, and pose usability challenges.

Therefore, contactless methods have gained attention as viable alternatives to traditional approaches, offering comfort, convenience, and reduced infection risk. Researchers have developed contactless medical instruments for various applications, including respiratory rate estimation by analyzing thoracic-abdominal region and face videos [4–8], heart rate estimation [9–13], tidal volume estimation [14–17], and thoracic-abdominal asynchrony (TAA) assessment [18–20]. Despite these advancements, a key indicator of ARD is still visually assessed by healthcare professionals. Chest retraction, considered an early sign of respiratory failure, is most commonly observed in infants and children but can also occur in patients with conditions such as asthma and pneumonia. Accurate and timely detection of chest retractions is essential, but reliance on visual examination poses challenges for consistent and continuous monitoring, leading to potential inaccuracies in assessment and outcomes.

In our previous work [21], we proposed an end-to-end ARD detection system that leveraged color temporal visual information in conjunction with advanced 3D deep convolutional neural networks, achieving high accuracy. However, this approach relied solely on color (RGB) temporal data, whose performance could be affected by variations in skin tone and lighting conditions. To overcome these limitations, we propose the use of multi-modality (RGB-D) temporal visual information for ARD detection. Compared to RGB data, RGB-D information provides additional depth insights that significantly enhance detection accuracy and model robustness. To effectively utilize this multi-modality information, we employ a two-stream model architecture combined with a late feature fusion scheme.

To sum up, this paper contributes to this field in the following ways:

1. We propose the use of multi-modality data to improve the performance of acute respiratory distress detection systems.
2. We introduce straightforward yet effective data pre-processing techniques to normalize the depth modality to ensure uniform scaling.
3. We investigate various feature fusion methods to effectively integrate information from both RGB and depth modality. Our experimental results demonstrate that simple feature fusion techniques are especially beneficial when working with limited data, resulting in significant improvements in detection performance.

The rest of this paper is structured as follows: Section 2 reviews relevant literature on current techniques for analyzing respiratory parameters and methods for multi-modality feature fusion. Section 3 provides an overview of the proposed model, detailing the pre-processing techniques, feature extraction module, and multi-modality feature fusion. Section 4 describes the database, implementation details, and presents the experimental results. Finally, Section 6 discusses the findings and provides concluding remarks.

2. Related Work

2.1. Methods for Respiratory Parameter Analysis

Methods for analyzing respiratory parameters are generally classified into two categories: contact-based and contactless approaches. Contact-based methods involve direct physical sensors attached to the body, such as respiratory inductance plethysmography (RIP) and pulse oximeters. In contrast, contactless methods employ non-invasive techniques, such as cameras or radar, which offer greater comfort and are particularly suitable

for newborns. These methods have garnered increasing interest due to their potential for improved functionality and integration with advancing technologies.

For example, Mateu et al. [22] used two color cameras to capture visual information and applied dense optical flow analysis to track motion for respiratory parameter estimation. Similarly, Rehouma et al. [17] utilized two 3D cameras to capture temporal point-cloud data, applying surface reconstruction techniques to accurately model the thoracoabdominal surface. They then calculated the volume for each frame and measured the respiratory rate through a volume–time graph. In another study, Rehouma et al. [18] proposed a method to assess thoracic-abdominal asynchronous motion using a single RGB-D camera, which calculates the 3D scene flow between consecutive frames to analyze motion. Additionally, V. Ottaviani et al. [20] developed a contactless method utilizing depth cameras to monitor infants' breathing patterns and thoracoabdominal asynchronous movements. Nawaz et al. [21] employed an RGB camera to capture the visual temporal information of patients, which was subsequently analyzed using 3D convolutional neural networks (CNNs). This approach aimed to non-invasively identify respiratory distress conditions by recognizing subtle visual cues associated with thoracoabdominal movements. It is important to note that only a limited number of studies have explored the ARD detection task through either contact-based or contactless methods.

2.2. Multi-Modality Fusion Techniques

Deep convolutional neural networks (DCNNs) are designed to capture data features. However, their performance can be influenced by variations in skin tone [23,24] and lighting conditions [25], particularly when trained on limited or biased RGB datasets that fail to adequately represent such diversity. In contrast, depth information remains consistent regardless of these factors, offering greater robustness, while depth data may lack the rich detail present in RGB images. It provides complementary information that can enhance overall performance when combined with RGB data. To fully leverage the strengths of both modalities, it is essential to fuse them into a comprehensive set of discriminative features.

Khalid et al. [26] proposed a multi-modal three-stream fusion network, drawing inspiration from the success of two-stream fusion networks [27,28]. This approach incorporates RGB spatial information, dense optical flow (temporal) data, and pose features to enhance model performance. Similarly, Islam et al. [29] introduced a multi-modal human activity recognition method that utilizes both RGB and depth temporal information. They employed a multi-modal feature fusion approach, specifically leveraging a self-attention mechanism to improve activity recognition accuracy.

Das et al. [30] designed an attention mechanism specifically to fuse spatial-temporal features with pose features, aiming to enhance the understanding of human actions. Joze et al. [31] developed the multi-modal transfer module (MMTM), a technique designed to progressively fuse features from both RGB and depth modalities, thereby improving model performance. Additionally, Hu et al. [32] utilized a bilinear pooling layer to effectively combine multi-modal features, further enhancing overall model efficacy.

Xu et al. [33] proposed a bilinear-pooling attention network to fuse RGB and skeleton features for action recognition tasks, showcasing the effectiveness of this fusion approach. Kini et al. [34] adopted an ensemble modeling strategy to leverage multi-modal information, achieving first place in the ICIAP-W 2023 challenge. Numerous other fusion [35–38] schemes have been proposed to effectively combine multi-modal information, highlighting the growing interest and research in this area.

3. Proposed Model

In this paper, we propose a two-stream network for detecting acute respiratory distress (ARD) by leveraging multi-modal data that incorporates both RGB and depth temporal visual (video) data. Our approach utilizes two identical 3D convolutional neural networks (CNNs) to independently extract spatiotemporal features from each modality, enabling a more comprehensive analysis of the visual cues associated with ARD conditions. By utilizing the strengths of both RGB and depth data, our network mitigates the limitations inherent in using only the RGB modality. These features are subsequently fused using a neural network, enabling the model to integrate complementary information and enhance overall ARD detection system performance. An overview of the proposed architecture is shown in Figure 1.

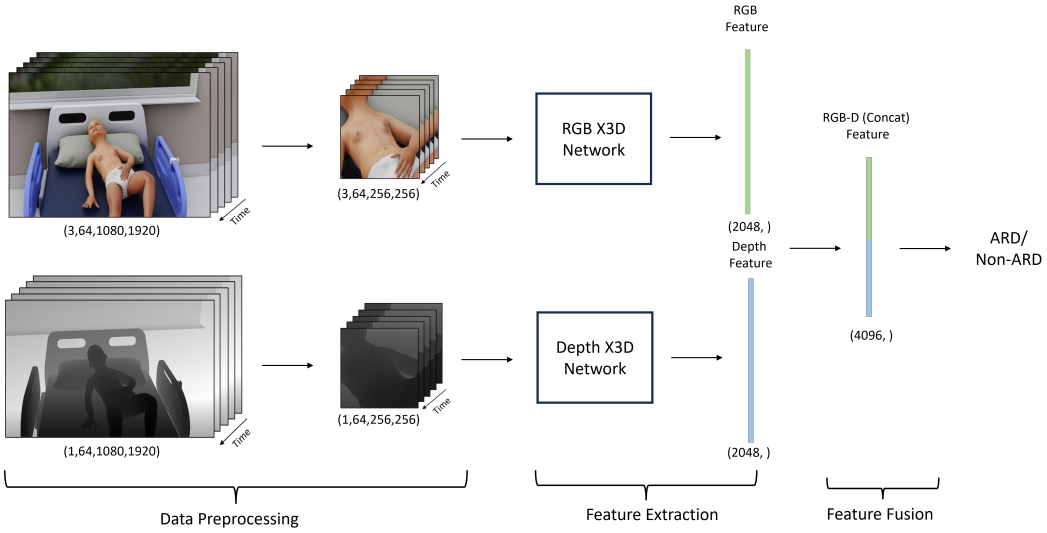


Figure 1. Illustration of the proposed network architecture for detecting acute respiratory distress, featuring the integration of RGB and depth temporal visual data through identical 3D convolutional neural networks.

In this section, we first formulate the problem and outline the data processing pipeline for both RGB and depth videos, describing the pre-processing steps implemented to prepare the data for analysis. Next, we describe the feature extraction strategy, where 3D convolutional neural networks are employed to capture the spatiotemporal characteristics of the input data. Lastly, we discuss the feature fusion techniques used to effectively combine the extracted features from both modalities, enhancing the model's overall performance.

3.1. Problem Formulation

The detection of acute respiratory distress (ARD) is defined as a video classification task, as the signs of retraction begin to appear at the start and continue throughout the inspiration cycle. To accurately detect ARD, our objective is to analyze the patient's video information over the entire inspiration cycle. A previous study [21] has demonstrated that a 6.4 s video clip is sufficient for accurate detection, as the respiratory cycle of an adult typically lasts up to 6.4 s. This duration ensures a high likelihood of capturing at least one full inspiration cycle, making it suitable for the ARD detection task.

3.2. Data Pre-Processing Module

For our experimental study, we use data collected from Sainte-Justine Hospital in Montreal, Canada. The data are captured using Microsoft Azure sensors, which simultaneously record RGB and depth information. The RGB data are captured using a 12-megapixel sensor, while the depth data are captured using a 1-megapixel sensor. The depth videos are

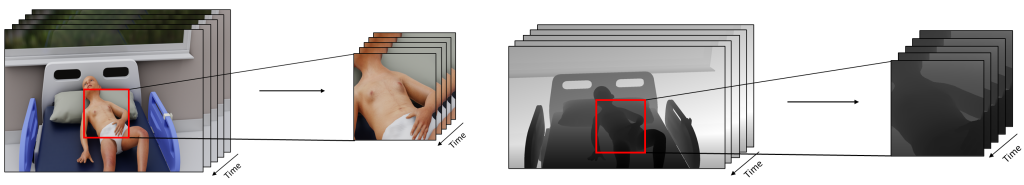
recorded at resolution of 512 × 512 in NFOV binned mode. In this mode, the sensor has an operational range of 0.50 to 5.46 m. The physical pixel size is approximately 0.0087 mm at 1 m. However, the physical pixel size varies depending on the distance to the object. RGB and depth videos were not spatially aligned and as they were recorded at different resolutions. For instance, the RGB videos have a resolution of 1080 × 1920 and a depth have 512 × 512.

To address this issue, we first align the RGB and depth videos using the Open3D library [39], which leverages the Azure Kinect Sensor SDK for alignment. Specifically, we align the depth videos to match the RGB videos' resolution. In addition, the collected data contain unnecessary background information that can negatively impact the performance of video analysis algorithms. This extraneous information can lead to overfitting, especially when working with limited data and high memory usage. To mitigate this issue and help our model focus solely on the relevant areas of the patients, we cropped both the RGB and depth videos to isolate these specific regions, as shown in Figure 2. This step is inspired by previous studies [21,33,40,41], which demonstrated that deep learning models trained on relevant regions of interest outperform those trained on full-frame data. Therefore, we have adopted a similar approach, extracting the thoracic-abdominal regions, where retraction signs typically appear.

Further, we spatially normalize the depth videos to a range of 0 to 1 for consistent scaling. We first remove outlier pixel values greater than 4000 (since the distance between the camera and the patient is not greater than that) by replacing them with zeros. Then, we compute the average distance of the thoracic-abdominal region by taking the mean of the non-zero pixels and selecting pixel values within a range of ±400 from this mean. Finally, the selected values are divided by 800. The pseudocode for the depth video normalization process is presented in Algorithm 1.

Algorithm 1: Pseudo code for depth video normalization process.

Input: Depth video (D)
Output: Normalized depth video (D_{norm})
foreach *Frame in depth video D* **do**
    Replace pixel values greater than 4000 with 0 (D_1);
    Compute the mean M of non-zero pixels of thoracic-abdominal region (D_1) ;
    Select pixels of (D_1) in the range $[M - 400, M + 400]$ and set 0 ;
    Scale selected pixel values by dividing by 800;
end
Return normalized depth video D_{norm};



(**a**) RGB (**b**) Depth

Figure 2. RGB-D videos' cropping: (**a**) RGB and (**b**) depth.

3.3. Feature Extraction Module

We use two X3D [42] (Expanding Architectures for Efficient Video Recognition) networks as feature extractors. X3D is a neural network architecture designed for video recognition tasks. It builds upon the 2D ConvNet architecture and progressively expands

it along multiple axes, including depth, width, resolution, and frame rate, to efficiently capture spatiotemporal features. By employing a series of lightweight 3D convolutional layers, X3D achieves high performance with fewer trainable parameters, which reduces computational resource requirements compared to traditional 3D-CNNs. This makes it particularly suitable for cases with limited data and real-time video analysis applications. We first train the two separate networks in an end-to-end manner for each modality using the ARD dataset. These models are trained to learn modality-specific features pertinent to the detection of ARD. After training, we use these networks as feature extractors in our model.

3.4. Feature Fusion Module

Numerous feature fusion techniques have been proposed, such as bilinear pooling [32] and self-attention networks [33]. However, these methods often involve additional fully connected layers with a large number of parameters, which can lead to overfitting, particularly when dealing with limited data. Considering the constraints (limited data) of our task, we chose to adopt a simpler approach. In this study, we employ a straightforward yet effective late fusion scheme based on feature concatenation, which demonstrates competitive results [33].

Models trained separately on the ARD dataset for the ARD task are then used to extract features by removing the classification layer. The extracted features are concatenated into a single 1D feature vector of size 4096, effectively combining the complementary information from both RGB and depth data. Figure 1 shows the feature fusion process. Finally, a simple single-layer neural network is trained to process the concatenated feature vector and make the final decision regarding the presence of ARD. This approach is characterized by its simplicity and effectiveness.

4. Experimental Analysis

4.1. Datasets

To evaluate the effectiveness of using multi-modalities for the ARD task, we conduct experiments on an ARD patient dataset. The dataset was collected at the Sainte-Justine Hospital Pediatric Intensive Care Unit (PICU) with approval from the Review Ethics Board (REB) (Ste-Justine REB number 2016-1242, approved on 31 March 2016) and parental consent.

Videos were recorded for each patient with a respiratory illness for a duration of 30 s. In total, we collected 210 videos in the PICU, with each video representing a unique patient. However, videos where the patient's torso region was covered, of poor quality, or with excessive noise were excluded. The remaining videos were labeled by two professionals using the Silverman scoring system [43], where the presence of at least one retraction indicated ARD. One professional labeled the data in real time during the recording process, and the second analyzed the videos to ensure information is captured effectively. Videos with labeling conflicts were also removed, resulting in a final dataset of 153 patients. Out of the 153 patients, 133 are aged 6 years or younger, with 63.16% exhibiting ARD, while the remaining 20 patients are older than 6 years, with 11.11% exhibiting ARD. The data distribution of ARD and Non-ARD patients categorized by age group, retraction type, and overall totals is presented in Figure 3. The figure on the left presents the ARD patients' statistics with respect to age. The figure in the middle presents the distribution of retraction signs in ARD patients. The figure on the right presents the overall class-wise data distribution. Of the 153 patients, 86 exhibit ARD, with signs of retractions distributed as follows: subcostal (74), intercostal (28), substernal (40), and suprasternal (16). The remaining 67 patients show no signs of chest retraction, indicating the absence of ARD.

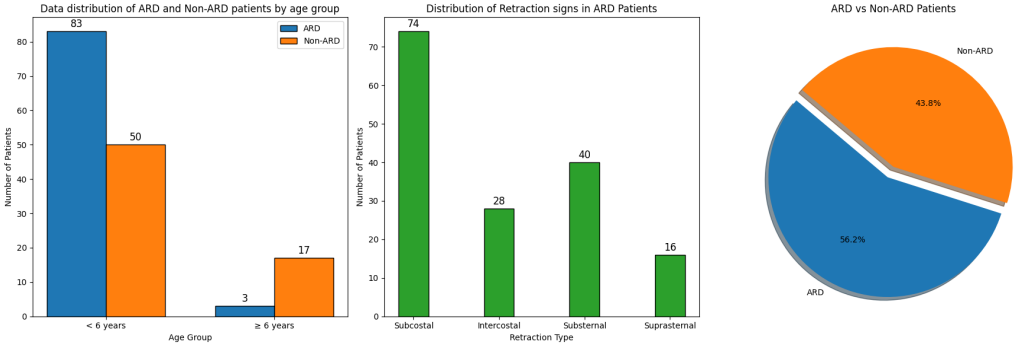


Figure 3. Data distribution of ARD and non-ARD patients categorized by age group, retraction type, and overall totals.

4.2. Implementation Details

We conduct all experiments using the PyTorch 2.5.1 framework on a NVIDIA Tesla V100-PCIE-32GB GPU. For model training and testing, we first split the data into training and validation sets using an iterative data splitting technique [44], based on the information of the chest retraction signs. The dataset of 153 patients (86 ARD and 67 non-ARD) is divided into 70% for training (60 ARD and 47 non-ARD) and 30% for testing (26 ARD and 20 non-ARD), ensuring balanced group allocation. The training dataset is further expanded by segmenting each video into 13 overlapping clips of 6.4 s, yielding a total of 1498 clips: 780 from ARD patients and 718 from non-ARD patients. We then spatially crop the videos to the shorter side to maintain the aspect ratio and resize them to 256 $\times$ 256 pixels. Additionally, we temporally sub-sample the videos to 10 frames per second (fps). We normalize the RGB and depth videos to a 0–1 pixel range. All data processing techniques are similar for both modalities, except for depth modality videos, which undergo spatial normalization (0–1), as described in Sections 3–3.2. We use stochastic gradient descent with a fixed learning rate of 0.0005 and a momentum of 0.9. The batch size is set to 64, using gradient accumulation techniques and binary cross-entropy loss is employed. We train the model for 40 epochs, saving the best checkpoints by monitoring the validation loss. During training, we apply temporal and spatial data augmentation techniques such as random spatial cropping to 224 $\times$ 224, temporal jittering, random rotation (±30 degrees), and horizontal and vertical flipping. During inference, we divide each video into four non-overlapping clips of 6.4 s, applying similar data pre-processing techniques as during training. The final prediction for each patient is determined by averaging the scores of the four clips.

4.3. Evaluation Metrics

To ensure a fair comparison, we maintain consistent training and testing configurations in the data splits. Additionally, we employed a five-fold cross-validation approach due to the limited size of the dataset. For the evaluation, we used standard metrics commonly used in classification tasks, including accuracy, precision, recall, and the F_1 score.

4.4. Ablation Study

To evaluate the effectiveness of multi-modal approaches, we first established baseline models for each modality. This step allowed us to establish a reference point for comparison before exploring the potential benefits of integrating multiple data modalities. Subsequently, we evaluated various feature fusion techniques, including early fusion through channel concatenation, late fusion methods such as feature concatenation, feature concatenation with a frozen backbone, and score averaging. We then compared the results to understand the contribution of multi-modal data to model performance.

4.4.1. Baseline

To establish baseline results, we evaluate three popular video analysis deep learning algorithms: X3D convolutional neural networks, channel-separated convolutional neural networks (CSNs) and R(2+1)D convolutional neural networks. We train all three algorithms independently on both RGB and depth modalities. Table 1 presents the experimental results of these three architectures. The results are reported for each evaluation metric as the minimum (min), average (avg), and maximum (max) score across five folds.

Table 1. Five-fold cross-validation results for three video analysis algorithms, X3D, CSN, and R(2+1)D, on RGB and depth modalities. Performance metrics (accuracy, precision, recall, and F_1 score) are reported as the minimum (min), average (avg), and maximum (max) scores across five folds.

Model		Accuracy			Precision			Recall			F_1 Score		
		Min	**Avg**	**Max**	**Min**	**Avg**	**Max**	**Min**	**Avg**	**Max**	**Min**	**Avg**	**Max**
X3D	**RGB**	0.783	0.822	0.870	0.818	0.872	0.95	0.72	0.777	0.833	0.783	0.821	0.864
	Depth	0.696	0.757	0.826	0.639	0.716	0.808	0.840	0.910	0.958	0.750	0.799	0.840
CSN	**RGB**	0.783	0.835	0.891	0.792	0.911	1.0	0.75	0.769	0.792	0.792	0.832	0.884
	Depth	0.565	0.665	0.739	0.593	0.668	0.750	0.640	0.745	0.875	0.615	0.701	0.750
R(2+1)D	**RGB**	0.717	0.796	0.826	0.7	0.793	0.84	0.792	0.835	0.875	0.764	0.812	0.857
	Depth	0.565	0.730	0.804	0.559	0.729	0.826	0.792	0.808	0.833	0.655	0.763	0.809

For the RGB modality, X3D achieves an average accuracy of 0.822, precision of 0.872, recall of 0.777, and an F_1 score of 0.821. For the depth modality, X3D attains an average accuracy of 0.757, precision of 0.716, recall of 0.910, and an F_1 score of 0.799. R(2+1)D also performs well on both RGB and depth modalities. It achieves an average accuracy of 0.796, precision of 0.793, recall of 0.835, and an F_1 score of 0.812 for the RGB modality. For the depth modality, R(2+1)D attains an average accuracy of 0.730, precision of 0.729, recall of 0.808, and an F_1 score of 0.763. CSN shows better performance on the RGB modality, achieving an average accuracy of 0.835, precision of 0.911, recall of 0.769, and an F_1 score of 0.832. For the depth modality, CSN achieves an average accuracy of 0.665, precision of 0.668, recall of 0.745, and an F_1 score of 0.701.

4.4.2. RGB-D Acute Respiratory Distress Detection

The experimental results in Table 1 show that the depth modality alone lacks sufficient information to capture chest retraction signs, which are crucial for ARD detection. Therefore, it is recommended to integrate the depth with RGB to enhance model robustness. To assess the effectiveness of multi-modality integration, we conduct experiments with two widely used fusion schemes: early fusion and late fusion. In the early fusion approach, the depth channel is integrated with the RGB channels, treating the depth data as a fourth channel, a method referred to as channel concatenation (CC). For the late fusion approach, we evaluate three variations: feature concatenation (FC), score averaging (SA), and feature concatenation with frozen base models for both modalities (FCF). The block diagram of the different types of multi-modality fusion schemes is presented in Figure 4.

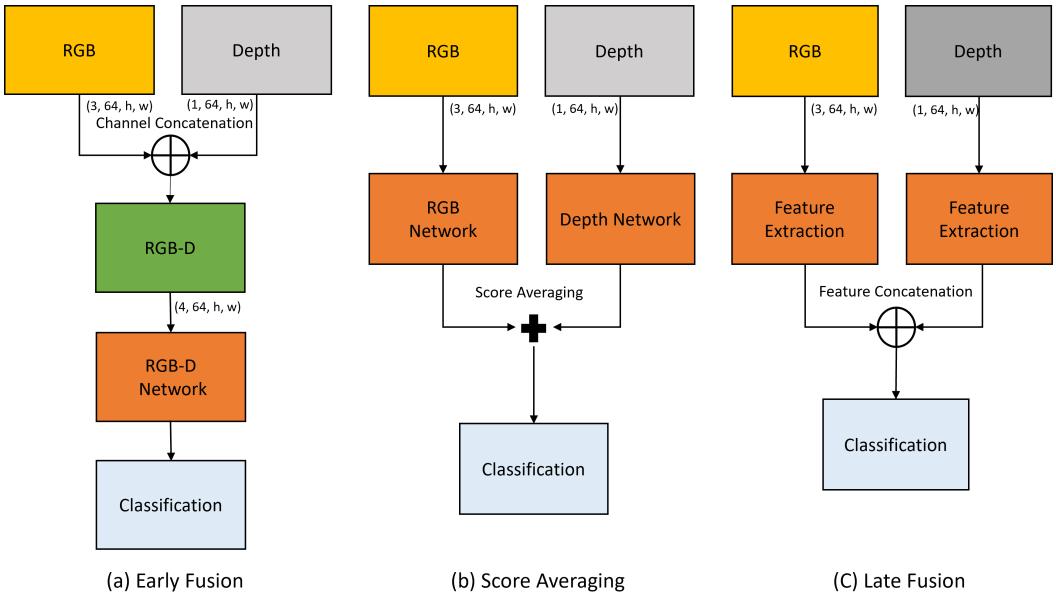


Figure 4. Block diagram illustrating the various types of multi-modality fusion schemes: (**a**) early fusion, where input modalities are combined at the input level; (**b**) score averaging, where individual modality predictions are averaged; (**c**) late fusion, where features are combined after independent processing of each modality w/o base-model freezing.

4.4.3. Early Fusion

In our first approach, we adapt a single deep learning-based video analysis algorithm to handle four-channel input by modifying the input and the first convolutional layer. Specifically, we use a pre-trained X3D model, expanding its first convolutional layer to accommodate the additional depth channel while maintaining the same number of filters and filter sizes. The layer weights are initialized by averaging the weights from the RGB model. The results of this channel concatenation (CC) fusion scheme are presented in the second row of Table 2. The results are reported for each evaluation metric as the minimum (min), average (avg), and maximum (max) score across five folds. The model achieved an average accuracy of 0.756, a precision of 0.762, a recall of 0.818, and an F_1 score of 0.788.

Table 2. Performance comparison of different feature fusion techniques across five folds (CC—channels concatenation, FCAT—feature concatenation, SA—score averaging and FCAT-F, feature concatenation with freezing base-model). Performance metrics, including accuracy, precision, recall, and F_1 score, are presented as the minimum (min), average (avg), and maximum (max) scores across the five folds.

Fusion Method	Accuracy			Precision			Recall			F_1 Score		
	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max
Baseline	0.783	0.822	0.870	0.818	0.872	0.95	0.72	0.777	0.833	0.783	0.821	0.864
CC	0.696	0.765	0.848	0.667	0.762	0.870	0.792	0.818	0.833	0.741	0.788	0.851
FCAT	0.804	0.830	0.848	0.800	0.821	0.840	0.833	0.868	0.917	0.816	0.843	0.863
SA	0.804	0.830	0.848	0.759	0.808	0.846	0.875	0.893	0.917	0.830	0.847	0.863
FCAT-F	0.804	0.852	0.913	0.808	0.867	0.917	0.760	0.852	0.917	0.809	0.858	0.917

4.4.4. Late Fusion

Due to the shortcomings of our initial approach, we adopt a late fusion strategy and evaluate three different schemes: feature concatenation (FCAT), score averaging (SA),

and feature concatenation with frozen base models (FCAT-F). In the FCAT approach, we employ a two-stream network to independently extract features from both RGB and depth modalities. These features are then concatenated at a later stage and used for classification. For the SA approach, we utilize two identical models to predict ARD condition scores for each modality, similar to the previous approach. These scores are aggregated and used for the final detection. Both FCAT and SA models are trained end-to-end. In contrast to the previous approaches, the FCAT-F method involves first training the models independently on each modality. Once trained, these models are used as feature extractors by removing their classification layers. The features from both modalities are then concatenated and passed to a single-layer neural network. During training, the weights of the base models, which were trained on the hospital dataset, are frozen, allowing only the fully connected single-layer neural network to be trained.

The experimental results of the late fusion schemes are presented in the last three rows of Table 2. The third row presents the results of the FCAT technique, which achieved an average accuracy of 0.830, a precision of 0.821, a recall of 0.868, and an F_1 score of 0.843, which is better than the RGB model. The fourth row shows the results of the SA technique, which achieved an average accuracy of 0.830, a precision of 0.808, and a recall of 0.893, The results of the FCAT-F technique are shown in the fifth row, demonstrating superior performance with an average accuracy of 0.852, a precision of 0.867, a recall of 0.852, and an F_1 score of 0.858.

4.4.5. Performance Analysis Across Age Groups

For this analysis, we group the dataset into two age groups: 1 (<6 years) and 2 ($\geq$6 years). As the dataset is biased toward the younger age group, similar trend in model performance is observed. The model exhibits strong performance in the first group (<6 years) with high accuracy (0.85) and precision (0.9047), while its performance in the second group ($\geq$6 years) is less favorable, with lower precision and all metrics. The imbalance between the groups likely influence the observed results. As there are only three positive examples in whole dataset for patients above 6 years old (2 for training and 1 for testing). The results presented in Table 3 represent the average performance across five-fold cross-validation. The average recall, precision, and true-positive rate (TPR) suggest that the model is unreliable for patients older than 6 years.

Table 3. Model performance across age groups (1: <6 years, 2: $\geq$6 years).

Age Group	Accuracy	Precision	Recall	TP_Rate	TN_Rate
1	0.788	0.863	0.767	0.767	0.820
2	0.744	0.166	0.400	0.400	0.796

5. Discussion

The experimental results highlight the importance of using multi-modality data for detecting ARD. Figure 5 presents the average accuracy, precision, recall, and F_1 score of the ARD detection system using different modality and different modality fusion schemes. It was found that the depth modality lacks the necessary information required for detecting chest retraction signs, a critical indicator of ARD. This limitation is primarily due to the low resolution of the depth camera (1 megapixel), which struggles to capture fine details, such as the subtle changes in lung pressure associated with chest retraction. Consequently, models relying on the depth modality face difficulties in learning crucial low-level, task-specific features. All three video analysis algorithms support the conclusion that the model struggles to perform with the depth modality alone (Table 1). However, networks using

the depth modality were able to make predictions based on the motion features caused by patient restlessness. However, these motion features are not discriminative, as they appeared in both non-ARD and ARD cases, limiting their value for distinguishing between the two conditions.

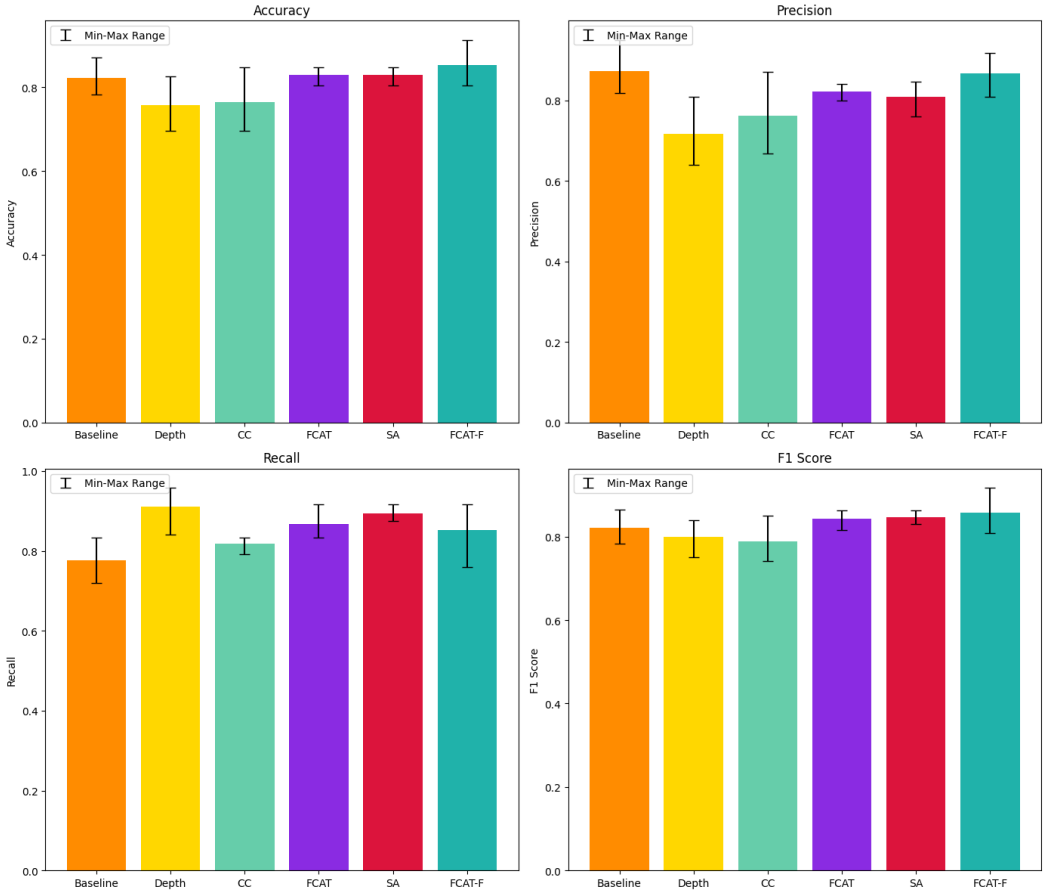


Figure 5. Performance comparison of ARD detection (X3D) system using different modality and different modality fusion schemes. The bars represent the average performance, with error bars indicating the min–max range of each metric across five folds.

Secondly, the experimental outcomes of integrating RGB-D data through early fusion scheme is even worse than using the RGB modality alone. A primary reason for the poor performance is the limited data size, which caused the model to overfit quickly, resulting in suboptimal performance. And, the re-initialization weights of the first convolutional layer of the network limited the potential benefits of transfer learning. In contrast, the integration of RGB-D information through late fusion schemes demonstrated significantly improved performance over single-modality approaches. Specifically, the FCAT-F approach emerged as the most effective fusion strategy in this study, achieving the highest average accuracy and F_1 score across all other fusion methods. This improvement is indicative of the benefits of independently training separate models for each modality. By doing so, each model is able to learn more distinct and task-specific features before combining them, leading to a more comprehensive and effective feature representation. This contrasts with end-to-end trained two-stream models (FCAT & SA), where feature learning may be less specialized. In summary, this study shows that using multi-modality information with an effective feature fusion scheme significantly improves ARD detection system performance.

6. Conclusions and Future Work

This study presented a two-stream multi-modal acute respiratory distress detection system utilizing 3D convolutional neural networks to analyze both RGB and depth data. The proposed system employs a late feature fusion scheme (feature concatenation) to integrate information from both modalities effectively. Experimental results demonstrate that the depth modality alone does not provide sufficient information for the ARD detection task. Furthermore, the results show that early fusion techniques are less effective for ARD detection, likely due to the limitations of the dataset size. In contrast, late fusion techniques, particularly the feature concatenation with freezing base models (FCAT-F) approach, substantially improve performance by effectively combining multi-modal information. The superior performance of FCAT-F underscores the advantages of leveraging pre-trained models and carefully integrating features from multiple sensors. However, the proposed method exhibits biased toward younger age groups (less than six years old), as only limited instances are available for patients aged more than 6 years.

For future work, we plan to explore pre-trained action recognition models specifically trained on RGB-D data, combined with advanced feature fusion techniques. In particular, we aim to investigate multi-level slow fusion and late fusion methods by initially training on large-scale datasets such as NTU RGB+D 120 and subsequently fine-tuning on our ARD dataset. This approach aims to leverage the rich features from larger datasets to address the challenges posed by limited data and enhance detection accuracy and robustness.

Author Contributions: Conceptualization, W.N. and R.N.; data curation, K.A. and P.J.; formal analysis, W.N. and K.A.; funding acquisition, P.J. and R.N.; methodology, W.N., P.J. and R.N.; project administration, P.J. and R.N.; software, W.N.; supervision, P.J. and R.N.; validation, W.N.; writing—original draft preparation, W.N.; writing—review and editing, K.A., P.J. and R.N. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada, and in part by the Fonds de Recherche du Québec—Santé (FRQS).

Institutional Review Board Statement: The study was conducted according to the guidelines of Research Centre of Sainte-Justine Hospital, QC, Canada and approved by the Review Ethic Board of Sainte-Justine Hospital (protocol code 2016-1242, approved on 31 March 2016).

Informed Consent Statement: Informed consent was obtained from all subjects involved in this study.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Edwards, M.O.; Kotecha, S.J.; Kotecha, S. Respiratory distress of the term newborn infant. *Paediatr. Respir. Rev.* **2013**, *14*, 29–37. [CrossRef] [PubMed]
2. Diamond, M.; Peniston, H.L.; Sanghavi, D.K.; Mahapatra, S. Acute respiratory distress syndrome. In *StatPearls*; StatPearls Publishing: Treasure Island, FL, USA, 2023.
3. Taussig, L.M.; Landau, L.I. *Pediatric Respiratory Medicine E-Book*; Elsevier Health Sciences: Amsterdam, The Netherlands, 2008.
4. Chen, M.; Zhu, Q.; Zhang, H.; Wu, M.; Wang, Q. Respiratory rate estimation from face videos. In Proceedings of the 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), Chicago, IL, USA, 19–22 May 2019; IEEE: New York, NY, USA, 2019; pp. 1–4.
5. Rehouma, H.; Noumeir, R.; Essouri, S.; Jouvet, P. Quantitative assessment of spontaneous breathing in children: Evaluation of a depth camera system. *IEEE Trans. Instrum. Meas.* **2019**, *69*, 4955–4967. [CrossRef]
6. Fiedler, M.A.; Rapczyński, M.; Al-Hamadi, A. Fusion-based approach for respiratory rate recognition from facial video images. *IEEE Access* **2020**, *8*, 130036–130047. [CrossRef]

7. Cheng, J.; Liu, R.; Li, J.; Song, R.; Liu, Y.; Chen, X. Motion-Robust Respiratory Rate Estimation from Camera Videos via Fusing Pixel Movement and Pixel Intensity Information. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 4008611. [CrossRef]
8. Fiedler, M.A.; Werner, P.; Rapczyński, M.; Al-Hamadi, A. Deep face segmentation for improved heart and respiratory rate estimation from videos. *J. Ambient Intell. Humaniz. Comput.* **2023**, *14*, 9383–9402. [CrossRef]
9. Gupta, P.; Bhowmick, B.; Pal, A. Accurate heart-rate estimation from face videos using quality-based fusion. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; IEEE: New York, NY, USA, 2017; pp. 4132–4136.
10. Pilz, C.S.; Zaunseder, S.; Krajewski, J.; Blazek, V. Local Group Invariance for Heart Rate Estimation From Face Videos in the Wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Salt Lake City, UT, USA, 18–23 June 2018.
11. Sabokrou, M.; Pourreza, M.; Li, X.; Fathy, M.; Zhao, G. Deep-hr: Fast heart rate estimation from face video under realistic conditions. *Expert Syst. Appl.* **2021**, *186*, 115596. [CrossRef]
12. Gao, H.; Wu, X.; Geng, J.; Lv, Y. Remote heart rate estimation by signal quality attention network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2122–2129.
13. Su, L.; Wang, Y.; Zhai, D.; Shi, Y.; Ding, Y.; Gao, G.; Li, Q.; Yu, M.; Wu, H. Spatiotemporal Sensitive Network for Non-Contact Heart Rate Prediction from Facial Videos. *Appl. Sci.* **2024**, *14*, 9551. [CrossRef]
14. Yuthong, A.; Duangsoithong, R.; Booranawong, A.; Chetpattananondh, K. Monitoring of volume of air in inhalation from Triflo using video processing. *IEEE Trans. Instrum. Meas.* **2019**, *69*, 4334–4347. [CrossRef]
15. Hurtado, D.E.; Chavez, J.A.; Mansilla, R.; Lopez, R.; Abusleme, A. Respiratory volume monitoring: A machine-learning approach to the non-invasive prediction of tidal volume and minute ventilation. *IEEE Access* **2020**, *8*, 227936–227944. [CrossRef]
16. Addison, P.S.; Smit, P.; Jacquel, D.; Addison, A.P.; Miller, C.; Kimm, G. Continuous non-contact respiratory rate and tidal volume monitoring using a Depth Sensing Camera. *J. Clin. Monit. Comput.* **2022**, *36*, 657–665. [CrossRef]
17. Rehouma, H.; Noumeir, R.; Bouachir, W.; Jouvet, P.; Essouri, S. 3D imaging system for respiratory monitoring in pediatric intensive care environment. *Comput. Med. Imaging Graph.* **2018**, *70*, 17–28. [CrossRef] [PubMed]
18. Rehouma, H.; Noumeir, R.; Masson, G.; Essouri, S.; Jouvet, P. Visualizing and quantifying thoraco-abdominal asynchrony in children from motion point clouds: A pilot study. *IEEE Access* **2019**, *7*, 163341–163357. [CrossRef]
19. Di Tocco, J.; Massaroni, C.; Bravi, M.; Miccinilli, S.; Sterzi, S.; Formica, D.; Schena, E. Evaluation of thoraco-abdominal asynchrony using conductive textiles. In Proceedings of the 2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Dubrovnik, Croatia, 25–28 May 2020; IEEE: New York, NY, USA, 2020; pp. 1–5.
20. Ottaviani, V.; Veneroni, C.; Dellaca, R.L.; Lavizzari, A.; Mosca, F.; Zannin, E. Contactless monitoring of breathing pattern and thoracoabdominal asynchronies in preterm infants using depth cameras: A feasibility study. *IEEE J. Transl. Eng. Health Med.* **2022**, *10*, 4900708. [CrossRef]
21. Nawaz, W.; Jouvet, P.; Noumeir, R. Automated Detection of Acute Respiratory Distress Using Temporal Visual Information. *IEEE Access* **2024**, *12*, 142071–142082. [CrossRef]
22. Mateu-Mateus, M.; Guede-Fernandez, F.; Angel Garcia-Gonzalez, M.; Ramos-Castro, J.J.; Fernández-Chimeno, M. Camera-based method for respiratory rhythm extraction from a lateral perspective. *IEEE Access* **2020**, *8*, 154924–154939. [CrossRef]
23. Merler, M.; Ratha, N.; Feris, R.S.; Smith, J.R. Diversity in faces. *arXiv* **2019**, arXiv:1901.10436.
24. Buolamwini, J.; Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the Conference on Fairness, Accountability and Transparency, PMLR, New York, NY, USA, 23–24 February 2018; pp. 77–91.
25. Adjabi, I.; Ouahabi, A.; Benzaoui, A.; Taleb-Ahmed, A. Past, present, and future of face recognition: A review. *Electronics* **2020**, *9*, 1188. [CrossRef]
26. Khalid, M.U.; Yu, J. Multi-Modal Three-Stream Network for Action Recognition. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 3210–3215. [CrossRef]
27. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 568–576.
28. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1933–1941.
29. Islam, M.M.; Iqbal, T. Hamlet: A hierarchical multimodal attention-based human activity recognition algorithm. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 25–29 October 2020; IEEE: New York, NY, USA, 2020; pp. 10285–10292.
30. Das, S.; Sharma, S.; Dai, R.; Bremond, F.; Thonnat, M. Vpn: Learning video-pose embedding for activities of daily living. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part IX 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 72–90.
31. Joze, H.R.V.; Shaban, A.; Iuzzolino, M.L.; Koishida, K. MMTM: Multimodal transfer module for CNN fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13289–13299.

32. Hu, J.F.; Zheng, W.S.; Pan, J.; Lai, J.; Zhang, J. Deep bilinear learning for rgb-d action recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 335–351.
33. Xu, W.; Wu, M.; Zhao, M.; Xia, T. Fusion of skeleton and RGB features for RGB-D human action recognition. *IEEE Sens. J.* **2021**, *21*, 19157–19164.
34. Kini, J.; Fleischer, S.; Dave, I.; Shah, M. Ensemble Modeling for Multimodal Visual Action Recognition. *arXiv* **2023**, arXiv:2308.05430.
35. Wang, X.; Hu, J.F.; Lai, J.H.; Zhang, J.; Zheng, W.S. Progressive teacher-student learning for early action prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3556–3565.
36. Popescu, A.C.; Mocanu, I.; Cramariuc, B. Fusion mechanisms for human activity recognition using automated machine learning. *IEEE Access* **2020**, *8*, 143996–144014. [CrossRef]
37. Keskes, O.; Noumeir, R. Vision-based fall detection using st-gcn. *IEEE Access* **2021**, *9*, 28224–28236. [CrossRef]
38. Khalid, N.; Ghadi, Y.Y.; Gochoo, M.; Jalal, A.; Kim, K. Semantic recognition of human-object interactions via Gaussian-based elliptical modeling and pixel-level labeling. *IEEE Access* **2021**, *9*, 111249–111266. [CrossRef]
39. Zhou, Q.Y.; Park, J.; Koltun, V. Open3D: A modern library for 3D data processing. *arXiv* **2018**, arXiv:1801.09847.
40. Tsou, Y.Y.; Lee, Y.A.; Hsu, C.T.; Chang, S.H. Siamese-rPPG network: Remote photoplethysmography signal estimation from face videos. In Proceedings of the 35th Annual ACM Symposium on Applied Computing, Online, 30 March–3 April 2020; pp. 2066–2073.
41. Ouzar, Y.; Djeldjli, D.; Bousefsaf, F.; Maaoui, C. X-iPPGNet: A novel one stage deep learning architecture based on depthwise separable convolutions for video-based pulse rate estimation. *Comput. Biol. Med.* **2023**, *154*, 106592. [CrossRef] [PubMed]
42. Feichtenhofer, C. X3d: Expanding architectures for efficient video recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 203–213.
43. Hedstrom, A.; Gove, N.; Mayock, D.; Batra, M. Performance of the Silverman Andersen Respiratory Severity Score in predicting PCO2 and respiratory support in newborns: A prospective cohort study. *J. Perinatol.* **2018**, *38*, 505–511. [CrossRef] [PubMed]
44. Szymański, P.; Kajdanowicz, T. A Network Perspective on Stratification of Multi-Label Data. In Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications, Skopje, Macedonia, 22 September 2017; Volume 74, pp. 22–35.