IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 July 2024, accepted 17 September 2024, date of publication 25 September 2024, date of current version 8 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3467266

RESEARCH ARTICLE

Automated Detection of Acute Respiratory Distress Using Temporal Visual Information

WAJAHAT NAWAZ[1], PHILIPPE JOUVET[2], AND RITA NOUMEIR[1], (Member, IEEE)
[1]Biomedical Information Processing Laboratory, École de Technologie Supérieure, University of Quebec, Montreal, QC H3C 1K3, Canada
[2]Research Center at CHU Sainte-Justine, University of Montreal, Montreal, QC H3T 1J4, Canada

Corresponding author: Wajahat Nawaz (wajahat.nawaz.1@ens.etsmtl.ca)

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

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Review Ethic Board of Sainte-Justine Hospital.

ABSTRACT The Pediatric Intensive Care Unit (PICU) receives critically ill patients with shortness of breath and poor body oxygenation. Various respiratory parameters, such as respiratory rate, oxygen saturation level, and heart rate, are continuously monitored to timely adapt their management. With the advancement in technology, measurements of most parameters are carried out by medical instruments. However, some crucial parameters are still measured via visual examination, particularly the assessment of chest deformation, which is vital in assessing acute respiratory distress (ARD) conditions. However, visual examination is subjective and intermittent, prone to human error, and challenging to monitor patients round the clock. This subjectivity becomes problematic, especially in areas with a shortage of specialists, such as remote locations, developing countries, or during pandemics. In this paper, we propose an automated acute respiratory distress condition detection system, to address challenges associated with visual examination. The proposed approach utilizes a high-definition camera to capture patient temporal visual information and employs advanced deep-learning models to detect ARD condition. In order to test the feasibility, we collected video data of 153 patients, including both with and without ARD in the PICU. As the deep learning models require substantial amounts of data, and collecting data in the medical domain, particularly in the PICU, poses challenges. To overcome data limited problem, we utilized the problem-specific information, opted transfer learning and data augmentation techniques. Additionally, we compute baseline results of various video analysis algorithms for ARD detection task. Experimental results illustrate that the deep learning base video analysis algorithms have the potential to automate the visual examination process for the ARD detection task, by achieving an accuracy of 0.82, precision of 0.80, recall of 0.89, and F_1 score of 0.84.

INDEX TERMS Acute respiratory distress, deep convolution neural networks, retraction signs, silver-man scoring, transfer learning, video classification.

I. INTRODUCTION

The primary objective of the respiratory system is to ensure effective gas exchange in the bloodstream through inhalation and exhalation process. During inhalation, the contraction of the diaphragm and intercostals muscles increases the volume of the thoracic cavity, resulting in decreased pressure within the lungs. This pressure difference causes air to flow from the atmosphere into the lungs. Simultaneously, oxygen is transferred to the bloodstream, and carbon dioxide is transported from the bloodstream into the lungs. Under normal conditions, the lungs provide oxygen to vital organs and remove carbon dioxide. However, in the case of lung injuries or viral infections, either an adequate amount of oxygen cannot reach the bloodstream, or carbon dioxide cannot be effectively removed. As a result, the brain activates

The associate editor coordinating the review of this manuscript and approving it for publication was Sandra Costanzo.

FIGURE 1. Examples of chest retraction signs and their potential locations.

accessory respiratory muscles to assist and maintain proper gas exchange. This condition is widely known as Acute Respiratory Distress (ARD). One of the most common reasons for an infant's admission to the pediatric intensive care unit (PICU) [1]. In severe cases, ARD has a high mortality rate, with about 40% of patient deaths resulting from the condition [2].

ARD is a life-threatening condition, as prolonged and excessive use of accessory muscles may progress to respiratory failure and subsequent cardiopulmonary arrest [3]. Hence, early-stage diagnosis of ARD is of paramount significance in the PICU, because prompt detection and diagnosis facilitate timely intervention and treatment, markedly enhancing patient survival rates and preventing severe damage to vital organs [4]. Patients with ARD exhibit a range of visual and auditory indicators, including an increased breathing rate, an agitated or frightened look, abdominal breathing, chest retractions, and wheezing or grunting sounds [5]. Among these indicators, the identification of chest retraction is a crucial, frequently utilized, non-invasive method for evaluating the severity of ARD, especially in the PICU [6]. Figure 1 depicts four types of chest retractions and their potential locations. These include intercostal retractions situated between the ribs, substernal retractions at the bottom of the sternum bone, suprasternal retractions above the sternum bone, and subcostal retractions below the rib margin. Retraction signs are categorized into two groups: mild and severe. Mild retraction signs are subtle and may require careful observation for accurate detection, relying significantly on the expertise of doctors or healthcare professionals.

Unfortunately, relying solely on visual examination to identify these signs introduces various challenges. This method requires a substantial healthcare workforce, demanding significant time and effort. It is also prone to human error, as the examiner's subjective interpretation can result in inconsistencies when detecting and quantifying retraction signs. Continuous visual monitoring of patients around the clock presents logistical challenges. On top of that, the subtle nature of mild retraction signs makes them challenging to discern with the naked eye, increasing the likelihood of oversight. These subjective limitations pose even more significant problems, particularly in remote areas, developing countries, and during pandemics, where the availability of healthcare professionals is already constrained. Researchers have tackled the aforementioned challenges by creating medical instruments for various purposes, including respiratory rate estimation [7], [8], [9], [10], [11], [12], sleep apnea event detection [13], [14], [15], [16], tidal volume estimation [17], [18], [19], and chest deformation assessment [20], [21], [22].

In this paper, we have presented a novel approach that mimics the visual examination procedure conducted by doctors. The overview of the proposed ARD detection system is depicted in Figure 2. Initially, a high-definition camera captures the patient's temporal visual information, ensuring the inclusion of at least one complete cycle of either inspiration or expiration. Subsequently, it extracts the potential region of interest encompassing the patient's torso, which is then fed into the ARD detection block. The ARD detection block employs advanced video analysis algorithms for decision-making. Our primary objective is to explore the feasibility of replacing the traditional visual examination conducted by doctors with an automated approach. Through extensive experimentation, we have identified that narrowing down the input information and leveraging state-of-the-art deep learning-based video analysis techniques allows for the successful automation of the visual examination process. These findings mark a significant advancement in the field, opening new avenues for more efficient and accurate ARD assessment. In summary, this paper makes the following contributions:

1) We present an end-to-end automated system for detecting ARD conditions, leveraging temporal visual information from patients with the aid of an advanced deep learning model.
2) We designed a mechanism to capture temporal visual information for the Pediatric Intensive Care Units.
3) We collected video data from a wide range of patients (from 0 to 18 years old) experiencing ARD and computed the baseline results of video analysis algorithms.
4) Additionally, we proposed solutions to address the limited data problem by utilizing problem-specific information, such as the selection of region of interest (ROI).

The rest of the paper is organized as follows. Section II presents the related work and identifies the research gap. Section III details the data acquisition and labeling process. Section IV outlines the proposed methodology, including data processing steps such as temporal and spatial region of interest extraction, and the ARD detection system architecture. Section V covers the implementation details, experimental results, cross-model evaluation, and qualitative analysis. Section VI, discusses the results and limitations of the work. Finally, Section VII concludes the paper and outlines potential future work.

II. LITERATURE REVIEW

Traditionally, the respiratory rate was estimated by counting the respiration cycles (inspiration and expiration) for one

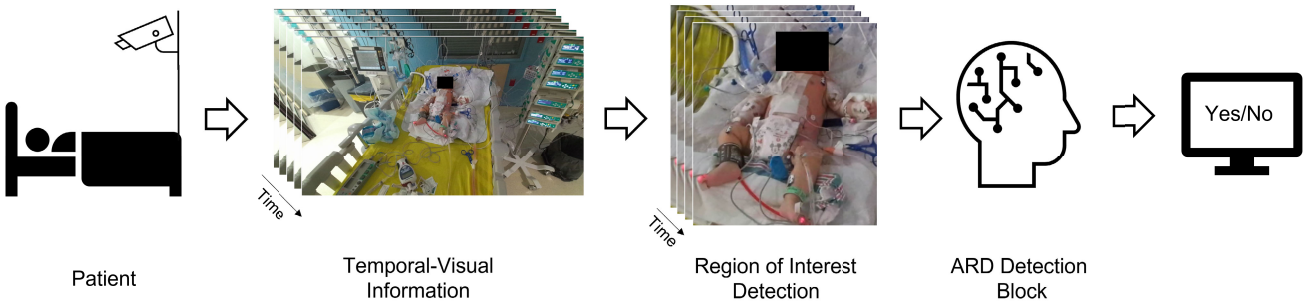

FIGURE 2. High level overview of our proposed acute respiratory distress detection system. 1). High-definition camera captures the temporal visual information and 2). ROI is extracted within the video then 3). fed into the ARD detection block.

minute. However, with the advancement of technology, Bai et al. [7] developed a contactless respiratory rate (RR) estimation device that can calculate the RR (whether too low or too fast) and generate an alarm if respiration stops for more than 10 seconds. Their system utilizes two color cameras to capture the patient's torso information and temporal differencing image processing to estimate RR. Likewise, Xia and Siochi [8] developed a contactless respiratory monitoring system employing a structured-light (SL) camera. The SL camera captures temporal torso color information and depth information, then estimates the RR by calculating the distance of the thoracic region over time. Benetazzo et al. [9] advanced the work of Xia and Siochi [8] by automating the thoracic region extraction step. They employed the *OPEN-AI* library to extract the thoracic region and compute its distance for each frame, and the distance-to-time graph peaks depict the RR.

In another study, Lee et al. [10] employed a Microwave Doppler Radar (MDR) sensor to estimate the distance-to-time graph for respiratory rate estimation. Mateu-Mateus et al. [23] designed a respiratory rate estimation device using an inexpensive camera. It captures the lateral perspective of the patient and estimates the motion between two consecutive frames using dense optical flow. Cheng et al. [12] introduced a motion-robust noncontact method for respiratory rate measurement, employing two-level fusion. This approach enhances RR estimation and improves reliability by considering the signal-to-noise ratio (SNR). Experimental findings indicate the method's superiority, offering potential advancements in video-based RR measurement.

Harte et al. [24] and Transue et al. [17] extended previous work to estimate another respiratory parameter, such as tidal volume. They employed a Microsoft Kinect camera to capture the point cloud information of the thoracic region and, with the help of a surface reconstruction algorithm, reconstructed the 3D surface. They then calculated the volume of each frame and used the volume-to-time graph and subtraction approach to measure the tidal volume. Similarly, Rehouma et al. [18] also proposed a contactless 3D imaging approach to estimate tidal volume and respiratory rate in the case of natural breathing. It employs 2× Time-of-Flight cameras to capture point cloud torso information from both lateral sides and register the point clouds into common world coordinates. After that, they employed a Poisson surface reconstruction [25] method to reconstruct the 3D surface of the thoracic region. The volume of the reconstructed surface was then measured using the Octree algorithm for each frame. Finally, the min-max subtraction technique was employed to measure the tidal volume. They tested their technique on an artificial mannequin with different settings (newborn, infant, and adult) and on two actual patient datasets [11].

Rehouma et al. [20] proposed a contactless 3D imaging-based system to recognize and quantify thoraco-abdominal asynchrony, a vital sign of respiratory distress in patients. Their system used a single RGB-D camera to capture torso information and calculate 3D scene flow [26] to segment the thoracic-abdominal region. They then used the Euclidean distance method to measure the thoracic and abdominal distances and plot the time vs. distance graph. They tested their methods on artificial mannequin simulations, not on actual patients. Di Tocco et al. [21] developed smart garments and examined thoracoabdominal asynchronies by analyzing time shifts between rib cage and abdomen movements. The smart garment comprises three elastic bands, each incorporating two conductive sensing elements that capture the motion of thoracic and abdominal regions. Ottaviani et al. [22] developed a contactless method for monitoring infants' breathing patterns and thoracoabdominal asynchronies using depth cameras. They employed depth cameras for precise depth analysis, which is essential for monitoring breathing patterns and thoracoabdominal asynchronies, making them a suitable choice for this specific medical application. They assessed their method for thoracoabdominal asynchronies on 12 patients with non-invasive respiratory support, evaluating its feasibility in clinical settings.

To our knowledge, no study has employed visual information for acute respiratory distress quantification. Most proposed methods address respiratory rate signals [7], [8], [9], [10], [11], [12] and tidal volume estimation [17], [18], [19], while very few discuss thoraco-abdominal asynchrony [20], [21], [22]. Additionally, these techniques do not report clinical experiments, deployment feasibility, or significant

constraints. In contrast to the aforementioned developments in the clinical field, our focus is to provide an end-to-end solution for ARD detection tasks under the constraints of PICU.

III. DATA ACQUISITION

Data acquisition for patients with Acute Respiratory Distress (ARD) is a challenging process. During the data collection phase, we faced two primary challenges: 1) patient body movements and occlusion due to hand movements and clothing, and 2) camera position. Doctors mitigate these problems by intervening during the visual examination to minimize their impact. However, to address these issues, we proposed recording the temporal-visual information for 30 seconds to predict whether the patient is experiencing ARD or not. This strategy aims to capture at least one static and unoccluded temporal region of the thoracic area for the ARD detection system. This longer-duration strategy allows us to mitigate the effects of the patient's body and hand movements and improve the accuracy of our approach. The second major challenge is the availability of camera positioning in the Pediatric Intensive Care Unit (PICU). After discussions with doctors, we identified four potential camera positions: the top right and left corners and the bottom right and left corners of the bed, as shown in Figure 3. However, we decided to dismiss the top left (a) and right (b) corner camera positions due to the occlusion of the suprasternal retraction sign from these two viewpoints. As a result, we selected the bottom camera positions for further experiments. The main position of the video acquisition tool was (d) bottom right, where there was no caregiver intervention, and sometimes (c). This decision was made in consultation with medical professionals to maximize the visibility of relevant visual information.

After review ethic board (REB) approval of the study (Ste-Justine REB number 2016-1242) and parent consent obtained, we employed a Microsoft Azure RGB-D sensor color camera (ultra-HD 12 megapixel RGB camera) to record patients' temporal-visual information at the CHU-Sainte-Justine Hospital's PICU. The recordings were primarily conducted during the patients' sleep periods to minimize unnecessary movements. However, it is worth noting that in many cases, patients showed noticeable movements involving their head, hands, and legs, ranging from slight to significant. During the data collection, we selected patients with respiratory conditions or potential candidates. This strategy ensured that our dataset encompassed a diverse range of cases related to respiratory conditions. By addressing the challenges, we aimed to create a comprehensive dataset that accurately represents various scenarios encountered in the PICU. The dataset contains individuals with diverse characteristics, including skin color and ethnicity, and a wide age range from 0 to 18. By incorporating this diversity, our dataset represents a broad population of patients in terms of skin color, gender, and age groups, enhancing the inclusivity and applicability of our proposed approach.

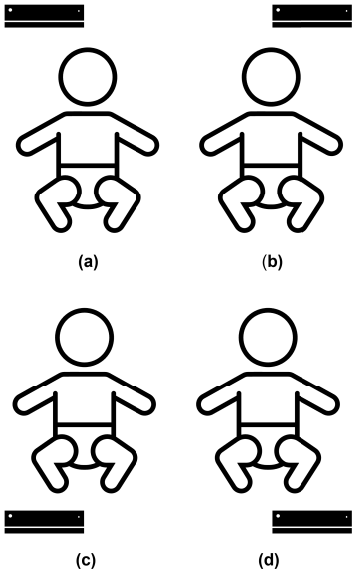

FIGURE 3. Available camera position in the pediatric intensive care units: top a). left and b). right, and bottom c). left and d). right.

A. DATA LABELING

In this study, 210 potential and respiratory distress patients participated. One video was recorded per patient. The videos were labeled by two professionals through visual and video analysis, utilizing the Silverman scoring method [27], considered the gold standard. The scoring method indicates mild respiratory distress in the presence of at least one mild retraction sign, while the presence of at least one severe retraction sign indicates severe respiratory distress. In the absence of any retraction sign, it indicates no respiratory distress. We asked two professionals to label the data in two different ways to demonstrate that videos can also be used for examination. Professional 1 labeled the videos during the recording process, while Professional 2 labeled the videos based on their analysis of the recorded footage. We compared the labels provided by the two professionals and removed cases with disagreements from the dataset to ensure its validity. As clinical evaluation is subjective, the scoring was done by two clinicians to achieve validated labeling.

Initially, the videos in the dataset were labeled into three classes: mild respiratory distress, severe respiratory distress, and no respiratory distress. For this study, we simplified the problem into a binary classification task. Among the 210 patients, 57 had their torsos fully covered by clothing, posing a challenge for detection and causing disagreements. Therefore, we excluded these cases, resulting in a total of 153 patients. Of these, 88 were labeled as having acute respiratory distress (ARD), while the remaining patients were labeled as non-ARD. Out of 88, 57 patients had mild respiratory distress, with the majority showing mild subcostal (50) and intercostal (25) retractions. In contrast, 31 patients were classified with severe respiratory distress, predominantly exhibiting severe subcostal (26) and substernal (7)

TABLE 1. Statistics of acute respiratory distress patients and associated retraction signs.

Acute Respiratory Distress	No. of patients	Retraction signs				
		Sub Costal	Inter Costal	Sub Sternal	Super Sternal	Supra Clavicular
Mild	57	50	25	33	14	5
Severe	31	26	4	7	2	0

retractions. Table 1 shows the detailed statistics of acute respiratory distress patients and the associated retraction signs.

IV. METHODOLOGY

Our approach for detecting Acute Respiratory Distress (ARD) involves a three-step process inspired by the visual examination procedures conducted by medical professionals. The first step entails using a high-resolution camera positioned on the bottom left or right side of the patient's bed in the pediatric intensive care unit. The camera, mounted on a stand at a 45-degree angle and placed 2 meters above the ground, captures temporal visual information. In the second stage, our system spatially identifies potential regions of interest in the acquired video. Lastly, we address ARD detection by framing it as a video classification problem. To ensure the self-contained nature and reproducibility of our work, we comprehensively describe the proposed approach in the following sections. We discuss the pre-possessing(spatial & temporal region of interest detection), and ARD detection. The pseudo code for the ARD detection system is presented in Algorithm 1.

Algorithm 1 Pseudo Code of Acute Respiratory Distress (ARD) Detection System

Input: Input video (V), 3D-CNN model (M) trained on ARD hospital data
Output: 1 **if** *ARD* exists, 0 otherwise
Initialize: Input video available for 6.4 seconds
Repeat
  Select the patient's Thoracic-abdominal region
  Spatially crop and resize the video to (256 × 256)
  Temporally sub-sample the video to 10 FPS
  Pass the pre-process video to the model (M)
  If score $\geq$ 0.5:
    Return 1 (patient has *ARD*)
  Else:
    Return 0 (patient has no *ARD*)
Until the input video is no longer available

A. TEMPORAL-SPATIAL REGION OF INTEREST EXTRACTION

This step is crucial as it enables the isolation of specific areas containing vital information relevant to Acute Respiratory Distress (ARD). To accomplish this, we employ a combination of spatial and temporal analysis techniques. Initially, a spatial extraction step is conducted to identify potential regions in the recorded video frames that are relevant to the ARD detection task. This involves segmenting the abdominal-thoracic regions, which are known to exhibit crucial visual cues for ARD detection. By isolating these specific regions, the system can concentrate on relevant areas, thereby discarding unnecessary information and enhancing both robustness and computational efficiency. Secondly, we temporally crop the videos to ensure they contain at least one complete inspiration/expiration cycle. This approach is undertaken because retraction signs become more prominent towards the end of the inspiration cycle.

TABLE 2. Respiratory rates for different age groups.

Age Group	Respiratory Rate (breaths per minute)
Infant (birth–1 year)	30-60
Toddler (1–3 years)	24–40
Preschooler (3–6 years)	22-34
School-age (6–12 years)	18-30
Adolescent (12–18 years)	12-16

1) SPATIAL SEGMENTATION

The recorded dataset consists of videos with a resolution of 1080 × 1920 and a frame rate of 30 FPS. Nonetheless, these high-resolution videos include unnecessary information, leading to computational inefficiency and overfitting issues, especially given the limited data. Unfortunately, existing video analysis techniques in the literature do not specifically address such high-resolution videos. Therefore, resizing the videos is necessary to improve computational efficiency and transform the data into a suitable format for further processing. However, simply resizing the videos without considering the content can result in the loss of essential information, particularly regarding the region of interest. For example, with the original video size of 1080 × 1920 and an ROI size of 256 × 256 (which varies from patient to patient), resizing the video frames to fit the data into the network can reduce the ROI size to 52 × 52. This resizing process leads to a significant loss of spatial information.

To overcome this problem, we propose spatially segmenting the videos by extracting the potential region of interest. In our case, the thoracic-abdominal region is identified as the potential ROI, as it primarily participates in respiratory activities and the retraction signs related to ARD predominantly appear in these areas. By extracting the potential ROI, we allow the network to concentrate exclusively on the relevant regions, helping it to learn more distinct and low-level features. Computational efficiency is also improved by narrowing down the videos, resulting in faster processing times and reduced power consumption. For now, we manually perform spatial segmentation on the videos to support our case studies.

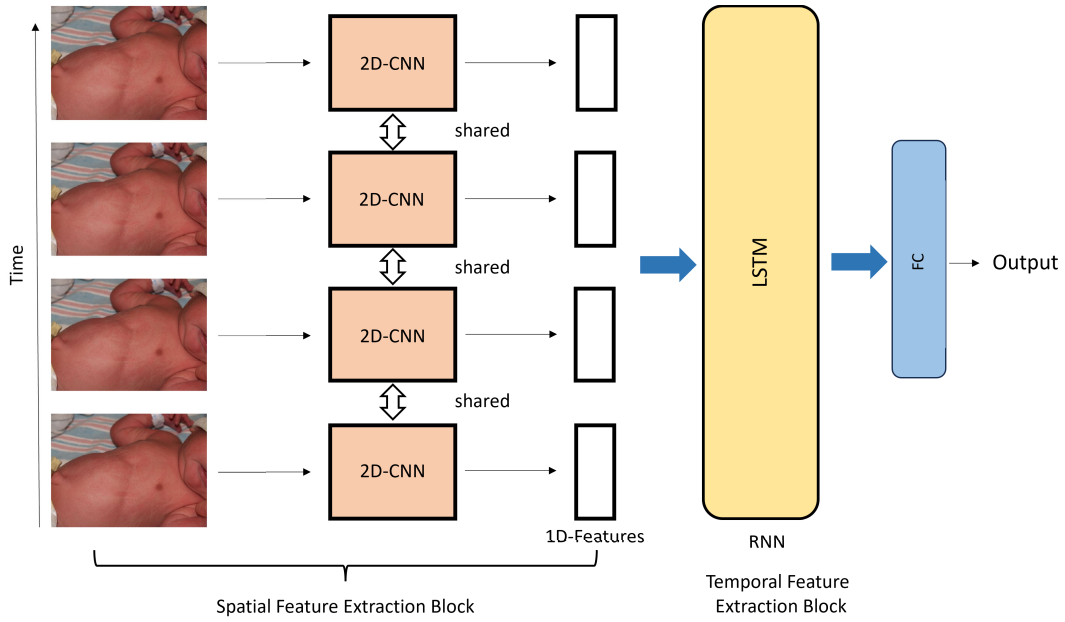

FIGURE 4. A 2D convolution neural network based respiratory distress (ARD) detection system. 1) Spatial Features Extraction Block: This block focuses on extracting spatial features from each video frame independently. 2) Temporal Feature Extraction Block: It extracts the temporal features that capture the relationships between frames. A recurrent neural network, specifically LSTM (Long Short-Term Memory), is employed to extract the temporal features.

2) TEMPORAL SEGMENTATION

During data collection, we recorded videos with a maximum duration of 30 seconds, capturing 30 frames per second (FPS), resulting in approximately 900 frames per video. However, processing such lengthy videos poses challenges in terms of computational efficiency and a higher risk of overfitting, particularly with a limited dataset size. No existing models in the literature are explicitly designed to handle such long-duration videos. Inspired by the temporal sliding window method commonly used in histopathology whole-slide image analysis [28], we applied a similar approach to our data by temporally dividing the videos into smaller segments. To determine the duration of these video segments, we considered respiratory rate (RR) statistics. Retraction signs, which are vital indicators of ARD, typically manifest from the start to the end of inspiration/expiration. The duration from the start to the end of inspiration/expiration is crucial for ARD detection. However, RR varies depending on the age and health condition of the patient. For example, infants have an RR of around 30-60 breaths per minute, which decreases to approximately 12-16 breaths per minute for adolescents. Table 2 displays RR per minute for different age groups. Extracting precise RR solely from RGB data is challenging, especially when the patient is making unnecessary movements. Considering RR states, an adolescent takes 6.4 seconds to complete one respiratory cycle and 3.2 seconds for inspiration/expiration. Therefore, a 3.2-second video clip is deemed ideal for our tasks. However, due to a lack of knowledge about the exact start of inspiration/expiration, we chose 4.8 and 6.4-second video segments for analysis. The longer duration video clip allows us to capture at least one start and end of inspiration/expiration while considering the variation in respiratory rates. By focusing on this segment, we can effectively analyze the temporal features associated with ARD while managing computational resources efficiently.

B. ACUTE RESPIRATORY DISTRESS DETECTION NETWORK

Once we have spatially and temporally cropped and preprocessed videos, they are fed into the ARD detection network. We framed our ARD detection problem as a video classification or action recognition task. Traditionally, research addressed action recognition problems using spatial-temporal handcrafted features [29] and optical flow [30] techniques. However, recent advancements in deep convolutional neural networks (CNNs) have revolutionized vision-related tasks, including image classification [31], object localization and segmentation [32], and action recognition [33]. Therefore, motivated by the success of deep CNN architectures in visual tasks, we leverage their capabilities to address the respiratory distress detection problem. By utilizing deep CNN architectures, we aim to automatically learn discriminative features from the preprocessed video data, enabling the network to detect respiratory distress effectively.

Deep learning-based algorithms for video classification or action recognition are categorized into two main categories: *2D-CNNs* combined with recurrent neural networks (RNNs) and *3D-CNNs*.

1) **2D-CNNs & RNNs**: Figure 4 shows the framework of a *2D-CNN + RNN* based ARD detection system.

This approach combines the strengths of *2D-CNNs* and *RNNs* [34] to capture spatial and temporal information. It employs *2D-CNNs* to extract spatial features from individual video frames. Then, the extracted spatial features are fed into a *RNNs*, such as LSTM (Long Short-Term Memory) [35] or GRU (gated recurrent unit) [36]. Initially, *RNNs* were designed to deal with time series data such text classification [37], sound classification [38]. Later on, researcher employed it for action recognition task [34], [39], [40] combining with *2D-CNNs*. *RNNs* maintain an internal hidden state that is updated over time, allowing the network to capture the temporal evolution of actions or movements. As the RNN processes the spatial features over time, it models the temporal dependencies between frames, effectively capturing the motion dynamics and temporal evolution of actions in the video. Additionally, *RNNs* uses shared weights to learn temporal features making it computationally efficient during interfering time. Combining the strengths of the *2D-CNNs* and *RNNs* allows the network to learn distinct spatial and temporal features from the videos.

2) **3D-CNNs**: Unlike the previous approaches, *3D-CNNs* [41] capture spatial and temporal information simultaneously. It employees 3D convolutional filters, which consider video data's width, height, and time dimensions. It allows the network to learn joint representations of appearance and motion features, comprehensively understanding the video content. By convolving 3D filters over the spatiotemporal volume of the video, *3D-CNNs* capture spatial and temporal correlations. The spatial dimension captures appearance-related features like shapes and textures, while the temporal dimension encodes motion-related features such as movement and dynamics. Learning these joint representations enables the network to recognize complex spatiotemporal patterns and distinguish different actions or activities based on their characteristics. *3D-CNNs* have demonstrated remarkable success in various video-related applications, including action recognition, video segmentation, and anomaly detection [41], [42], [43], [44]. Figure 5 shows 3D convolutions neural networks based ARD detection system.

V. EXPERIMENTAL RESULTS

We initiated experiments to examine the impact of spatial segmentation, various frame sampling rates (2,3,4,5,6) commonly used in action recognition task and temporal segmentation (3.2,4.8 and 6.4 seconds). Then, we experimented various types of deep learning based video-classification algorithms.

A. IMPLEMENTATION DETAILS

For model training and testing, we divide 30-second-long patient videos into smaller video clips, as discussed in Section IV-A2. Each video is segmented into, for example, 6.4-second clips consisting of 192 frames. We temporally sub-sample the videos at a rate of 2 (15 FPS), as per standard practices. Then, we spatially crop the videos to the shorter side of the frames to maintain the aspect ratio and resize them to $T \times 3 \times 256 \times 256$, where T is the number of frames. Stochastic gradient descent with an initial learning rate of 0.0005 and a momentum of 0.9 is utilized to mitigate the risk of converging to local minima. The optimization process employs a binary cross-entropy loss function and a batch size of 64, using gradient accumulation techniques.

To address small data and overfitting problems, several strategies were implemented. Firstly, the training videos were temporally divided into smaller chunks to increase the data size. Secondly, problem-specific crops were applied to the video to focus on the region of interest (ROI), specifically the torso region. Thirdly, we chose to implement transfer learning techniques rather than training the model from scratch, drawing inspiration from the encouraging outcomes reported in several studies that utilized smaller datasets. Additionally, online data augmentation techniques, including flipping, random cropping, rotating, temporally jittering, and early stopping were used.

B. DATA PREPROCESSING AND EVALUATION METRICS

Primarily, we spatially crop the videos, centering on the patients (whole body), as a prepossessing step for all experiments. We split 153 videos (one 30-seconds video per patient) into two disjoint sets (training - 70% and validation - 30%): T_1, composed of 107 videos, is the training set, and T_2, composed of 43 videos, is the validation set for fine-tuning model parameters. An iterative splitting method [45], based on retraction signs information, is used to equally distribute instances of each class into the training and validation sets. We artificially increase the training data size by temporally splitting videos (14 clips per video), resulting in a data size of 1498 clips. While during testing, we used four non-overlapping clips 6.4 seconds per video, to compute the average classification scores. However, in case of 3.2 and 4.8 seconds clips, we select 8 and 6 clips per video clips per video. Evaluation metrics include accuracy, precision, recall, and F_1 score were used to evaluate the performance of model. We employ three-fold cross-validation test at the patient level to ensure a reliable evaluation of the proposed approach's performance.

C. PRELIMINARIES RESULTS

In the experiments, we established baseline results and proposed solutions to enhance model performance. We set an initial benchmark by evaluating 3D-CNN-based video analysis algorithms on our ARD dataset. To assess accuracy, we employed the channel-separated 3D convolutional network (CSN) [42], recognized for its superior performance in human action recognition (HAR) tasks on the Kinetics-400 dataset. We fine-tuned the $CSN - R101$ originally trained

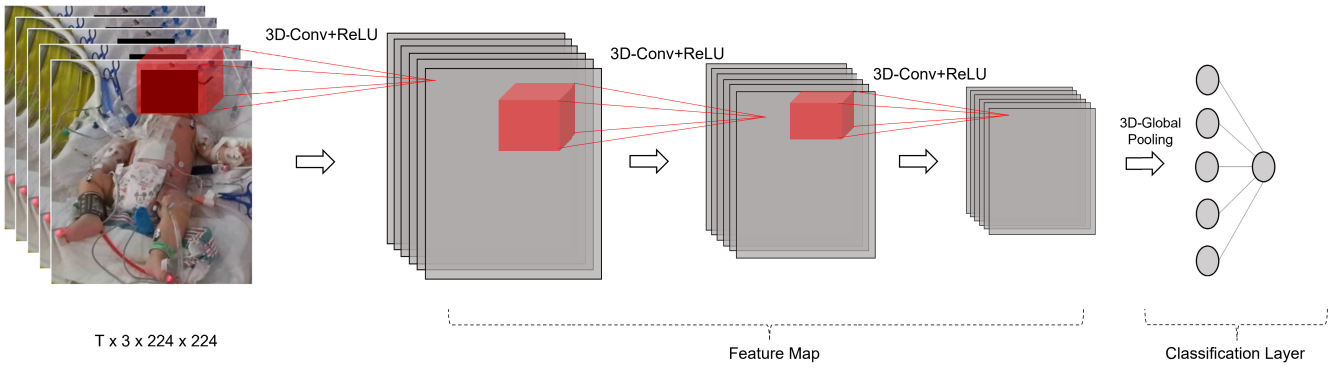

FIGURE 5. 3D Convolution Neural Network Based Acute Respiratory Distress detection network.

TABLE 3. Experimental results of with and without torso selection.

Torso Selection	Accuracy	Precision	Recall	F_1 Score
No	0.725	0.747	0.724	0.733
Yes	**0.812**	**0.809**	**0.849**	**0.828**

on the Kinetics-400 HAR dataset on the ARD dataset by modifying the classification head rather than training the model from scratch.

1) BASELINE

Table 3 shows the results of the deep learning-based ARD detection network on spatially cropped (full patient) videos with a clip duration of 6.4 seconds and a frame sampling rate of 2 (15 FPS). Our experimental results indicate that the model trained on full patient videos suffers from severe over-fitting problems. The model achieves an accuracy of 0.725, precision of 0.747, recall of 0.724, and an F_1 score of 0.733. This is due to the limited data, which causes the model to memorize high-level (background) information, such as external respiratory support devices, and fail to generalize during testing. Additionally, we experimented with and without spatial resizing and obtained similar performance.

2) SPATIAL SEGMENTATION

Table 3 presents the results of the deep learning-based ARD detection network with and without torso selection. The experimental results indicate that selecting the ROI significantly helps the model learn more distinct and relevant features compared to not selecting the torso, thereby enhancing the model's performance. The model achieves an accuracy of 0.812, precision of 0.809, recall of 0.849, and an F_1 score of 0.828 with torso selection. In contrast, without torso (ROI) cropping, the model's performance is noticeably lower, with an accuracy of 0.725, precision of 0.747, recall of 0.724, and an F_1 score of 0.733. This indicates that focusing on the region of interest, especially in the case of limited data, helps prevent overfitting and assists the model in learning distinct low-level features. In our subsequent experiments, we utilized data that had undergone torso selection.

TABLE 4. Experimental results of different frame sampling rates.

Frame Sampling Rate	Accuracy	Precision	Recall	F_1 Score	Computational Time(sec)
2	0.812	0.809	0.849	0.828	0.288
3	**0.819**	**0.798**	**0.891**	**0.840**	0.266
4	0.815	0.786	0.891	0.835	0.226
5	0.790	0.770	0.876	0.818	0.199
6	0.768	0.741	0.877	0.802	0.185

3) FRAME SAMPLING RATE

The optimal frame sampling rate (step size) depends on the specific information the model aims to detect [46]. In our ARD dataset, which involves newborn infants exhibiting varying respiratory rates (30 - 60 breaths per minute), careful selection of the frame sampling rate is crucial to avoid the loss of critical information. To explore this, we conducted experiments using five different frame sampling rates: 2, 3, 4, 5, and 6, which are commonly used in action recognition tasks. Throughout these experiments, we maintained a fixed 6.4-second video clip length and trained and tested the model with different frame sampling rates. The experimental results, presented in Table 4, show the model performance for each frame sampling rate setting and their computational cost for one clip. The results indicate that a frame sampling rate of 3 achieves the highest accuracy (0.819), precision (0.798), recall (0.891), and F_1 (0.840). However, as the frame

TABLE 5. Experimental results of different clip duration.

Clip Length (sec)	Accuracy	Precision	Recall	F_1 Score	Computational Time (Sec)
3.2	0.732	0.776	0.7	0.733	0.185
4.8	0.789	0.7610	0.891	0.819	0.226
6.4	**0.819**	**0.798**	**0.891**	**0.840**	0.266

TABLE 6. Experimental results of various video analysis algorithms and their computational time.

Model type		Configurations	Accuracy (min, avg, max)			Precision (min, avg, max)			Recall (min, avg, max)			F_1 Score (min, avg, max)			Computational time (sec)
2D-CNNs+LSTM	ResNet-50	(Learning rate:0.001, optimizer: SGD, momentum=0.90)	0.761	0.775	0.783	0.750	0.786	0.818	0.750	0.794	0.840	0.783	0.789	0.792	0.667
3D-CNNs	R(2+1)D	(Learning rate:0.0005, optimizer: SGD, momentum=0.90)	0.717	0.775	0.826	0.70	0.775	0.833	0.792	0.822	0.84	0.864	0.796	0.833	0.279
	S3D		0.717	0.754	0.783	0.731	0.77	0.818	0.75	0.767	0.792	0.745	0.768	0.783	0.254
	SlowFast R101		0.717	0.761	0.848	0.678	0.744	0.840	0.80	0.85	0.875	0.755	0.792	0.857	0.266
	X3D		0.783	0.804	0.847	0.769	**0.833**	0.905	0.76	0.795	0.833	0.792	0.812	0.844	0.245
	CSN-R101		0.761	**0.819**	0.870	0.733	0.798	0.875	0.875	**0.891**	0.917	0.80	**0.840**	0.875	0.266
	SWIM-VIT		0.717	0.812	0.891	0.688	0.821	0.913	0.792	0.849	0.88	0.772	0.831	0.894	0.79

sampling rate increases beyond 3 (larger step size), there is a gradual degradation in performance. The decrease in performance with higher sampling rates is likely due to the loss of essential temporal details for respiratory distress. This leads to the generation of false positive samples during training, causing the model to focus on irrelevant features. In contrast, low frame sampling rates (smaller step size) may introduce noise and add unnecessary information, making the model overly sensitive to small changes and thus less generalized.

4) TEMPORAL SEGMENTATION

In this experiment, we investigate the impact of varying the duration of video clips on the training and testing of the ARD detection model's performance. For this experiment, we use a fixed frame sampling rate of 3 (10 FPS). Table 5 presents the experimental results of the ARD detection model using different video clip durations and their corresponding computational time per clip. The clip lengths considered are 3.2 seconds, 4.8 seconds, and 6.4 seconds, as discussed in Section IV-A2. The results for a video clip duration of 3.2 seconds are presented in the first row of Table 5. The model achieved an accuracy of 0.732. Due to the absence of precise information regarding the start and end times of inspiration/expiration cycles, this limitation causes the training data loader to generate false positive examples, leading the model to learn irrelevant information.

However, as the clip duration increases, the model's accuracy improves from 0.732 to 0.789 and 0.819, as shown in the second and third rows of Table 5. This improvement is because longer video clips provide the model with more comprehensive temporal information, enabling it to better capture distinct and relevant features. The 6.4-second clip achieved the highest performance with the best F_1 score of 0.840, emphasizing the importance of considering a longer duration for effective ARD detection. In summary, a clip duration of 6.4 seconds appears to be the most effective for achieving optimal performance in detecting respiratory distress based on the presented experimental results.

D. EXPERIMENTAL RESULTS OF VIDEO ANALYSIS ALGORITHM

In this study, we explore two main types of video analysis algorithms: $2D-CNNs$ with LSTM and $3D-CNNs$ for acute respiratory distress (ARD) detection. We conduct experiments using a 6.4-second video clip and a frame sampling rate of 3 (10 FPS) because it outperforms other settings. The rest of the configuration details are shown in Table 6. For the first type of model, we use $ResNet-50$ trained on the ImageNet dataset and train the LSTM layer from scratch on the ARD dataset. On the other hand, we use 3D-CNN models trained on the action recognition video dataset and fine-tune them on the hospital ARD database by changing the classification head. Additionally, we use data augmentation techniques to enhance the dataset and an early stopping function to improve performance and avoid overfitting. Furthermore, we run a 3-fold cross-validation test to assess the generalizability of the models, and the average score across the folds is used to evaluate the performance of various models. The results are summarized in Table 6, providing insights into the performance of the video analysis algorithm on our hospital's ARD database. The table shows the minimum, average, and maximum scores of each model across the 3-fold cross-validation. $2D-CNNs+LSTM$, employing ResNet-50 for spatial feature extraction and a single-layer LSTM for temporal information extraction, achieve an accuracy of 0.775. They demonstrate a reasonable balance between precision (0.786) and recall (0.794), with an F_1 score of 0.789.

In the evaluation of 3D-CNN models for Acute Respiratory Distress (ARD) detection, various architectures are explored, such as $R(2+1)D-R50$, $SlowFast-R101$, $X3D$, $CSN-R101$, and $SWIM-VIT$. The $R(2+1)D$, which uses the $ResNet-50$ architecture, shows reasonable performance in ARD

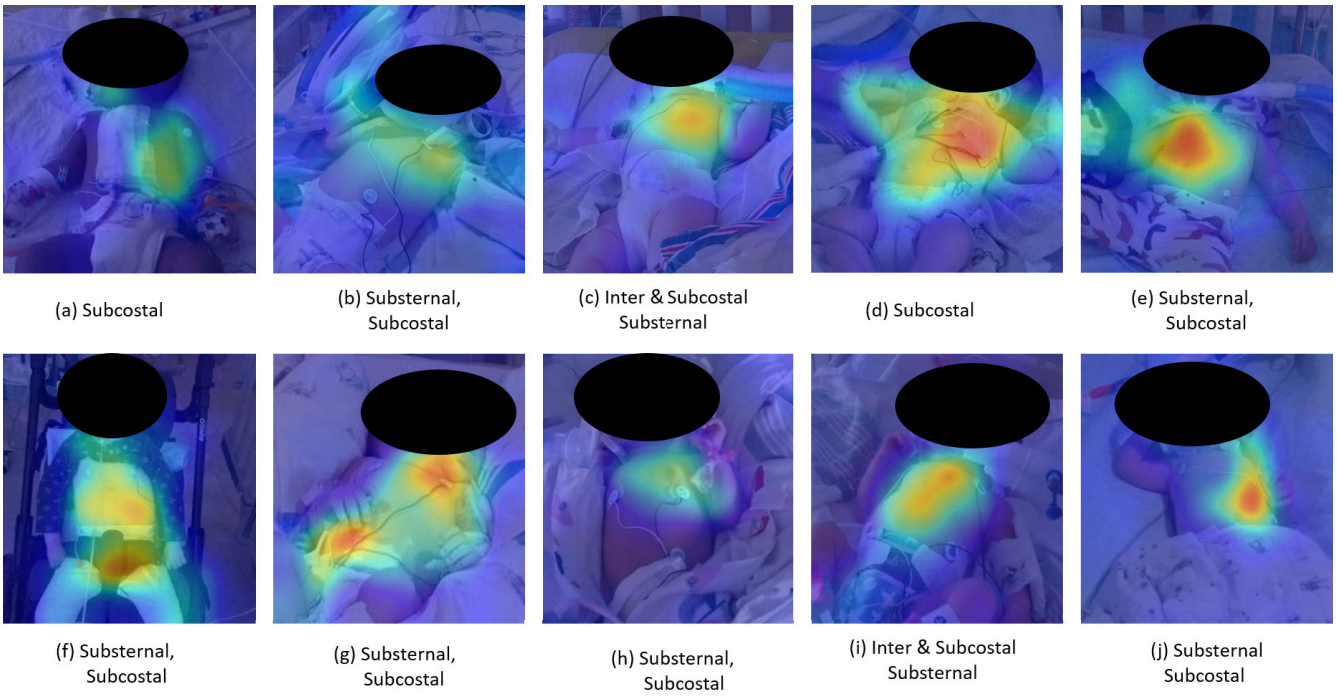

FIGURE 6. Qualitative results of an acute respiratory distress detection model using class activation maps (CAM) reveal that the model has learned problem-specific features, concentrating exclusively on the torso region of the patients. Nevertheless, in some cases such as (a), (b), (f), and (g), it also learns irrelevant features, focusing on non-interested regions.

detection, achieving an accuracy of 0.775 and demonstrating balanced precision and recall at 0.775 and 0.822, respectively. The $S3D$ model achieves an average accuracy of 0.75, precision of 0.77, recall of 0.77, and an F_1 Score of 0.77. The $SlowFast - ResNet - 101$ architecture shows moderate performance with an accuracy of 0.761. Notably, it displays effective recall at 0.85, showcasing its ability to correctly identify positive instances. The $X3D$ model achieves an accuracy of 0.804, precision of 0.833, recall of 0.80, and an F_1 score of 0.812. The $SWIM - VIT$ model achieves an accuracy of 0.812, precision of 0.821, recall of 0.849, and an F_1 Score of 0.831. The $CSN - R101$ model demonstrates superior performance compared to other models, with an accuracy of 0.819, a precision of 0.798, a recall of 0.891, and an F_1 score of 0.840. However, the performance variation between different models is due to the design and complexity of the model. The models such as $R(2+1)D$, $Slow - Fast - R101$, and $S3D$ are designed with a frame sampling rate of 5. As a result, they learn the large motion features. On the other hand, $X3D$, $CSN - R101$, and $SWIM - VIT$ are designed with a sampling rate of 2. Therefore, they are able to learn the small spatial-temporal features effectively. Because of this, they show optimal performance on the ARD hospital dataset. Additionally, the $3D - CNNs$ trained on an action recognition dataset naturally capture the temporal information, leading to optimal performance compared to $2D - CNNs + LSTM$. In terms of computational cost and evaluation metrics, the $CSN - R101$ model outperforms other models.

E. QUALITATIVE ANALYSIS

Figure 6 presents the qualitative results of an acute respiratory distress detection model using class activation maps (CAM). The findings reveal that the model has learned problem-specific features, concentrating on the thoracic-abdominal (torso) region of the patients, such as the rib cage and the area between the abdominal and rib cage. Remarkably, the model has also identified other important features, such as the agitated and frightened look (facial features) and restlessness (motion), which are additional symptoms of acute respiratory distress conditions. This suggests that the model has the potential to capture broader cues related to ARD beyond the specific regions of interest initially targeted. However, in some cases, such as (a), (b), (f), and (g), the model has also partially focused on irrelevant features, examining non-relevant regions that are not associated with ARD. These findings show that the deep learning model has the potential to automate acute respiratory distress detection.

VI. DISCUSSIONS

The experimental results emphasize the crucial role of carefully selecting data pre-processing techniques to improve the overall performance of acute respiratory distress (ARD) detection. The first experiment shows that focusing on the thoracic-abdominal area only when analyzing videos significantly boosts the model's performance, aiding in preventing overfitting, particularly when dealing with limited data. In contrast, the model trained on full patient videos suffers from severe overfitting problems. This is due to limited

data causing the model to memorize high-level (background) information, such as external respiratory support devices. Moreover, the exploration of different frame rates highlights the influence of frame sampling rates on the model's performance in the ARD detection task. A higher frame sampling rate results in the loss of essential temporal details for respiratory distress detection, leading to the generation of false positive samples during training and causing the model to focus on irrelevant features. Therefore, the optimal frame sampling rate is critical for effective respiratory rate detection. Additionally, the choice of the right video clip duration is crucial for effective ARD detection. Due to the lack of precise information about the exact start and end times of inspiration/expiration cycles, the training data loader generates false positive examples, leading the model to learn irrelevant features. On the other hand, longer video clips provide the model with more comprehensive temporal information during training and testing, enabling it to better capture distinct and relevant features. The $3D-CNNs$, especially $CSN-R101$ and $SWIM-VIT$, outperform $2D-CNNs+LSTM$ and other $3D-CNNs$ models, leveraging their ability to capture both large and small temporal features. However, our proposed approach faces challenges when patients make movements due to coughing and crying. These substantial movements make it challenging to accurately assess retraction signs even through visual examination. In summary, our model is sensitive to patient movements.

VII. CONCLUSION & FUTURE WORK

Acute respiratory distress is a life-threatening condition caused by lung diseases or viral infections. Traditional ARD detection methods are subjective, prone to human error, labor-intensive, and challenging for continuous 24/7 monitoring. To address these challenges, we have developed an innovative automated acute respiratory distress detection system using deep convolutional neural networks. We have demonstrated that state-of-the-art deep convolutional neural networks can effectively automate ARD detection tasks. Our proposed system overcomes the limitations of visual examination procedures and intermittent monitoring. If validated under clinical conditions, this method could help alleviate the shortage of medical specialists in remote areas, developing countries, and during pandemics. As part of future work, we plan to gather more data and automate the detection of spatial and temporal regions of interest, which play a significant role in the ARD detection model. We also intend to expand our scope beyond ARD detection to include the identification and quantification of retraction signs to assist doctors in a more effective manner. To obtain access to the data, please reach out to Philippe Jouvet. Note that specific institutional review board rules will apply.

REFERENCES

[1] M. O. Edwards, S. J. Kotecha, and S. Kotecha, "Respiratory distress of the term newborn infant," *Paediatric Respiratory Rev.*, vol. 14, no. 1, pp. 29–37, Mar. 2013.

[2] H. F. Ramji, M. Hafiz, H. H. Altaq, S. T. Hussain, and F. Chaudry, "Acute respiratory distress syndrome; A review of recent updates and a glance into the future," *Diagnostics*, vol. 13, no. 9, p. 1528, Apr. 2023.

[3] V. K. Jaeger, D. Lebrecht, A. G. Nicholson, A. Wells, H. Bhayani, A. Gazdhar, M. Tamm, N. Venhoff, T. Geiser, and U. A. Walker, "Mitochondrial DNA mutations and respiratory chain dysfunction in idiopathic and connective tissue disease-related lung fibrosis," *Sci. Rep.*, vol. 9, no. 1, p. 5500, Apr. 2019.

[4] V. Mirabile. (2023). *Respiratory Failure*. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK526127/

[5] National Heart, Lung, and Blood Institute (NHLBI). (2022). *Acute Respiratory Distress Syndrome Symptoms*. Accessed: Apr. 16, 2024. [Online]. Available: https://www.nhlbi.nih.gov/health/ards/symptoms

[6] E. D. McCollum and A. S. Ginsburg, "Outpatient management of children with world health organization chest indrawing pneumonia: Implementation risks and proposed solutions," *Clin. Infectious Diseases*, vol. 65, no. 9, pp. 1560–1564, Oct. 2017.

[7] Y.-W. Bai, W.-T. Li, and Y.-W. Chen, "Design and implementation of an embedded monitor system for detection of a patient's breath by double webcams in the dark," in *Proc. 12th IEEE Int. Conf. e-Health Netw., Appl. Services*, Jul. 2010, pp. 93–98.

[8] J. Xia and R. A. Siochi, "A real-time respiratory motion monitoring system using KINECT: Proof of concept," *Med. Phys.*, vol. 39, no. 5, pp. 2682–2685, May 2012.

[9] F. Benetazzo, A. Freddi, A. Monteriù, and S. Longhi, "Respiratory rate detection algorithm based on RGB-D camera: Theoretical background and experimental results," *Healthcare Technol. Lett.*, vol. 1, no. 3, pp. 81–86, Sep. 2014.

[10] Y. S. Lee, P. N. Pathirana, R. J. Evans, and C. L. Steinfort, "Noncontact detection and analysis of respiratory function using microwave Doppler radar," *J. Sensors*, vol. 2015, pp. 1–13, Jan. 2015.

[11] H. Rehouma, R. Noumeir, S. Essouri, and P. Jouvet, "Quantitative assessment of spontaneous breathing in children: Evaluation of a depth camera system," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 7, pp. 4955–4967, Jul. 2020.

[12] J. Cheng, R. Liu, J. Li, R. Song, Y. Liu, and X. Chen, "Motion-robust respiratory rate estimation from camera videos via fusing pixel movement and pixel intensity information," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–11, 2023.

[13] K. Feng, H. Qin, S. Wu, W. Pan, and G. Liu, "A sleep apnea detection method based on unsupervised feature learning and single-lead electrocardiogram," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2020.

[14] Q. Shen, H. Qin, K. Wei, and G. Liu, "Multiscale deep neural network for obstructive sleep apnea detection using RR interval from single-lead ECG signal," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[15] M. Bahrami and M. Forouzanfar, "Sleep apnea detection from single-lead ECG: A comprehensive analysis of machine learning and deep learning algorithms," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022.

[16] X. Chen, W. Ma, W. Gao, and X. Fan, "BAFNet: Bottleneck attention based fusion network for sleep apnea detection," *IEEE J. Biomed. Health Informat.*, vol. 28, no. 5, pp. 2473–2484, May 2024.

[17] S. Transue, P. Nguyen, T. Vu, and M.-H. Choi, "Real-time tidal volume estimation using iso-surface reconstruction," in *Proc. IEEE 1st Int. Conf. Connected Health: Appl., Syst. Eng. Technol. (CHASE)*, Jun. 2016, pp. 209–218.

[18] H. Rehouma, R. Noumeir, W. Bouachir, P. Jouvet, and S. Essouri, "3D imaging system for respiratory monitoring in pediatric intensive care environment," *Computerized Med. Imag. Graph.*, vol. 70, pp. 17–28, Dec. 2018.

[19] A. Yuthong, R. Duangsoithong, A. Booranawong, and K. Chetpattananondh, "Monitoring of volume of air in inhalation from triflo using video processing," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 7, pp. 4334–4347, Jul. 2020.

[20] H. Rehouma, R. Noumeir, G. Masson, S. Essouri, and P. Jouvet, "Visualizing and quantifying thoraco-abdominal asynchrony in children from motion point clouds: A pilot study," *IEEE Access*, vol. 7, pp. 163341–163357, 2019.

[21] J. Di Tocco, C. Massaroni, M. Bravi, S. Miccinilli, S. Sterzi, D. Formica, and E. Schena, "Evaluation of thoraco-abdominal asynchrony using conductive textiles," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (IMTC)*, May 2020, pp. 1–5.

[22] V. Ottaviani, C. Veneroni, R. L. Dellaca', A. Lavizzari, F. Mosca, and E. Zannin, "Contactless monitoring of breathing pattern and thoracoabdominal asynchronies in preterm infants using depth cameras: A feasibility study," *IEEE J. Transl. Eng. Health Med.*, vol. 10, pp. 1–8, 2022.
[23] M. Mateu-Mateus, F. Guede-Fernández, M. á. García-González, J. J. Ramos-Castro, and M. Fernández-Chimeno, "Camera-based method for respiratory rhythm extraction from a lateral perspective," *IEEE Access*, vol. 8, pp. 154924–154939, 2020.
[24] J. M. Harte, C. K. Golby, J. Acosta, E. F. Nash, E. Kiraci, M. A. Williams, T. N. Arvanitis, and B. Naidu, "Chest wall motion analysis in healthy volunteers and adults with cystic fibrosis using a novel Kinect-based motion tracking system," *Med. Biol. Eng. Comput.*, vol. 54, no. 11, pp. 1631–1640, Nov. 2016.
[25] M. Kazhdan, "Poisson surface reconstruction," in *Proc. 4th Eurographics Symp. Geometry Process.*, 2006, pp. 61–70.
[26] M. Jaimez, C. Kerl, J. Gonzalez-Jimenez, and D. Cremers, "Fast odometry and scene flow from RGB-D cameras based on geometric clustering," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 3992–3999.
[27] W. A. Silverman and D. H. Andersen, "A controlled clinical trial of effects of water mist on obstructive respiratory signs, death rate and necropsy findings among premature infants," *Pediatrics*, vol. 17, no. 1, pp. 1–10, 1956.
[28] Z. Zhu, L. Yu, W. Wu, R. Yu, D. Zhang, and L. Wang, "MuRCL: Multi-instance reinforcement contrastive learning for whole slide image classification," *IEEE Trans. Med. Imag.*, vol. 42, no. 5, pp. 1337–1348, May 2023.
[29] A. Klaeser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. Brit. Mach. Vis. Conf.*, 2008, pp. 1–275.
[30] U. Mahbub, H. Imtiaz, and Md. A. Rahman Ahad, "An optical flow based approach for action recognition," in *Proc. 14th Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dec. 2011, pp. 646–651.
[31] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," 2016, *arXiv:1602.07360*.
[32] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[33] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.
[34] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4694–4702.
[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[36] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (GRU) neural networks," in *Proc. IEEE 60th Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2017, pp. 1597–1600.
[37] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, Apr. 2019.
[38] M. Bubashait and N. Hewahi, "Urban sound classification using DNN, CNN & LSTM a comparative approach," in *Proc. Int. Conf. Innov. Intell. Informat., Comput., Technol. (ICT)*, Sep. 2021, pp. 46–50.
[39] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional LSTM with CNN features," *IEEE Access*, vol. 6, pp. 1155–1166, 2018.
[40] R. S. Kiziltepe, J. Q. Gan, and J. J. Escobar, "A novel keyframe extraction method for video classification using deep neural networks," *Neural Comput. Appl.*, vol. 35, no. 34, pp. 24513–24524, Dec. 2023.
[41] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.
[42] D. Tran, H. Wang, M. Feiszli, and L. Torresani, "Video classification with channel-separated convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5551–5560.
[43] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 305–321.
[44] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6201–6210.
[45] P. Szymanski and T. Kajdanowicz, "A network perspective on stratification of multi-label data," in *Proc. 1st Int. Workshop Learn. Imbalanced Domains, Theory Appl.*, vol. 74, P. B. Luís Torgo and N. Moniz, Eds., Sep. 2017, pp. 22–35.
[46] A. Diba, M. Fayyaz, V. Sharma, A. H. Karami, M. M. Arzani, R. Yousefzadeh, and L. Van Gool, "Temporal 3D convnets: New architecture and transfer learning for video classification," 2017, *arXiv:1711.08200*.

WAJAHAT NAWAZ received the B.Eng. degree in electrical (telecommunication) engineering from Government College University Faisalabad, Pakistan, and the M.S. degree in electrical engineering specializing in digital signal and system processing from the National University of Science and Technology (NUST). He is currently pursuing the Ph.D. degree in applied engineering with École de Technologie Supérieure (ETS), Canada, with a focus on machine learning for biomedical informatics. Previously, he was a Research Associate with the Intelligent Machine Laboratory, ITU, Pakistan, with a focus on a Facebook-funded project for real-time malaria detection. He was a Software Developer and a Machine Learning Engineer with Bahria University and NUST, Islamabad, Pakistan.

PHILIPPE JOUVET received the M.D. degree from Paris V University, Paris, France, in 1989, the M.D. degree in pediatrics and the M.D. degree in intensive care from Paris V University, in 1989 and 1990, respectively, and the Ph.D. degree in pathophysiology of human nutrition and metabolism from Paris VII University, Paris, in 2001. He joined the Pediatric Intensive Care Unit of Sainte-Justine Hospital, University of Montreal, Montreal, QC, Canada, in 2004. He is currently the Deputy Director of the Research Center and the Scientific Director of the Health Technology Assessment Unit, Sainte-Justine Hospital, University of Montreal. He conducts a research program on computerized decision support systems for health providers. His research program is supported by several grants from the Sainte-Justine Hospital, Quebec Ministry of Health, FRQS, Canadian Institutes of Health Research (CIHR), and the Natural Sciences and Engineering Research Council (NSERC). He has published more than 160 articles in peer-reviewed journals. He gave more than 120 lectures at national and international congresses. He has a salary award for research from Quebec Public Research Agency (FRQS).

RITA NOUMEIR (Member, IEEE) received the master's and Ph.D. degrees in biomedical engineering from École Polytechnique of Montreal. She is currently a Full Professor with the Department of Electrical Engineering, École de Technologie Supérieure (ETS), Montreal. She has extensively worked in healthcare information technology and image processing. She has also provided consulting services in large-scale software architecture, healthcare interoperability, workflow analysis, and technology assessment for several international software and medical companies, including Canada Health. Her main research interest includes applying artificial intelligence methods to create decision support systems.