ASSESSMENT OF THE ACUTE RESPIRATORY DISTRESS USING A DEPTH CAMERA

by

Wajahat NAWAZ

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, DECEMBER 2, 2025

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mrs. Rita Noumeir, Thesis supervisor
Department of Electrical Engineering, École de Technologie Supérieure

Mr. Philippe Jouvet, Thesis Co-Supervisor
Pediatric Intensivist - Ste. Justine Hospital Montréal, Université de Montréal

Mr. Simon Drouin, Chair, Board of Examiners
Department of Software Engineering and IT, École de Technologie Supérieure

Mr. Mohamad Forouzanfar, Member of the Jury
Department of Systems Engineering, École de Technologie Supérieure

Mr. James Green, External Examiner
Department of Systems and Computer Engineering, Carleton University

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON NOVEMBER 21, 2025

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

My journey toward a Ph.D. has been both challenging and deeply rewarding, and I am profoundly grateful for the unwavering support and encouragement I have received from my supervisors, colleagues, friends, and family. Although words can hardly express my sincere gratitude, I would like to take this opportunity to acknowledge their invaluable contributions to both my academic and personal growth.

First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Professor Rita Noumeir, for her invaluable guidance, insightful comments, and continuous support throughout my Ph.D. journey. Over the past few years, she has inspired me not only with her exceptional knowledge and research expertise but also with her commitment to academic excellence. I feel truly fortunate to have had her as my advisor. Thank you for your immense patience, encouragement, and belief in my work. I also thank my co-supervisors, Professor Phillipe Jouvet, for his valuable comments and feedback on my research manuscripts. I would also like to thank my Ph.D. assessment jury for their constructive feedback and support.

I am profoundly grateful to my fellow lab mates and friends for their unwavering support, insightful discussions, and camaraderie that sustained me through challenging times. In particular, I wish to thank Thanh-Dung LE for helping me maintain optimism during the toughest moments, and Toufik Bouras , whose friendship and thoughtful advice reminded me to enjoy life beyond the research lab. I am especially thankful to Mario Francesco for his constant humor and well-timed memes that added laughter when it was needed most. Finally, I wish to acknowledge Srinivasan Ramachandran for his steadfast support through every challenge.

I would also like to thank the entire Clinical Decision Support System (CDSS) team at CHU Sainte-Justine Hospital, in particular Kevin Albert and Edem Donatien Tiassou, for their valuable collaboration, technical assistance, and continued support throughout this research.

Finally, I wish to express my profound gratitude to my parents, wife, and siblings for their unwavering love, support, and faith in me. Their strength and understanding have been the cornerstone of this journey and have made this achievement possible.

ÉVALUATION DE LA DÉTRESSE RESPIRATOIRE AIGUË À L'AIDE D'UNE CAMÉRA DE PROFONDEUR

Wajahat NAWAZ

RÉSUMÉ

La détresse respiratoire aiguë constitue une phase précoce de l'insuffisance respiratoire caractérisée par une altération sévère des échanges gazeux, entraînant une oxygénation artérielle inadéquate et/ou une élimination insuffisante du dioxyde de carbone. En l'absence de traitement rapide, elle peut rapidement évoluer vers une insuffisance respiratoire. Cette urgence clinique se manifeste par divers signes observables reflétant les mécanismes compensatoires de l'organisme pour maintenir une oxygénation adéquate. Les manifestations cliniques courantes comprennent la tachypnée, le battement des ailes du nez, le geignement expiratoire, l'asynchronisme thoraco-abdominal et le tirage (rétractions thoraciques). Parmi ces indicateurs, les signes de tirage constituent des marqueurs hautement spécifiques et sensibles, car ils indiquent directement une augmentation du travail respiratoire. La présence de ces signes constitue une urgence médicale nécessitant une intervention clinique rapide. La détection prompte et précise de ces signes critiques est primordiale pour initier les interventions thérapeutiques appropriées et prévenir l'insuffisance respiratoire.

La pratique clinique actuelle repose principalement sur l'évaluation visuelle, un processus où les professionnels de la santé observent physiquement les patients au chevet pour évaluer la sévérité de la détresse respiratoire par l'identification des signes de tirage. L'évaluation visuelle offre plusieurs avantages, notamment son caractère non invasif, l'obtention de résultats immédiats et l'absence d'équipement spécialisé. Cependant, cette approche de surveillance manuelle et intermittente souffre de variabilité inter-observateurs et nécessite des ressources importantes, exigeant une supervision experte continue. Ces limitations sont particulièrement prononcées dans les contextes à ressources limitées et les scénarios pandémiques où les ressources cliniques sont sous tension.

Cette thèse présente un système de détection de la détresse respiratoire aiguë (DRA) sans contact basé sur l'intelligence artificielle, qui pallie les limites de l'examen visuel en automatisant le processus d'évaluation. Le système proposé exploite la technologie de caméra RGB-D (couleur et profondeur) pour capturer des informations visuelles et temporelles de la paroi thoracique du patient de manière non invasive et continue. Le système utilise en outre des modèles d'apprentissage profond pour localiser avec précision les régions de la paroi thoracique et segmenter des fenêtres temporelles cliniquement significatives, tout en éliminant efficacement les artefacts de mouvement. Des algorithmes avancés d'analyse vidéo extraient ensuite des caractéristiques spatio-temporelles discriminantes des flux de données multimodales raffinés pour l'identification automatisée de la détresse respiratoire.

Cette thèse apporte trois contributions clés à travers des études interconnectées. Premièrement, nous évaluons diverses architectures d'analyse vidéo basées sur l'apprentissage profond pour la détection de la DRA dans des contextes de données cliniques limitées. Notre évaluation

révèle que les ensembles de données cliniques réels présentent des biais spatiaux inhérents. Pour relever ce défi, nous proposons un cadre de sélection spatio-temporelle. Une évaluation systématique démontre que les régions cliniquement pertinentes et la longueur appropriée de la fenêtre temporelle sont critiques pour une détection précise et efficace en termes de calcul. Une analyse plus approfondie révèle que les modèles effectuant un sous-échantillonnage temporel parallèlement à l'extraction de caractéristiques spatiales démontrent des performances supérieures par rapport aux architectures qui conservent l'information temporelle complète. Le système de DRA proposé, exploitant le modèle CSN-R101, atteint une exactitude de 82%, une précision de 80%, un rappel de 89% et un score F_1 de 84%.

Deuxièmement, nous étudions la fusion de données multimodales pour améliorer la précision de détection. Nous établissons d'abord que l'information de profondeur seule est insuffisante pour une détection robuste de la DRA. Par la suite, nous démontrons que la fusion tardive des caractéristiques des modalités RGB et de profondeur surpasse substantiellement les approches unimodales, atteignant une exactitude de 85,2%, une précision de 86,7%, un rappel de 85,2% et un score F_1 de 85,8%, améliorant significativement les résultats du RGB seul (exactitude de 82,2%, précision de 87,2%, rappel de 77,7%, score F_1 de 82,1%). Ces résultats démontrent que bien que la profondeur seule soit inadéquate, elle fournit des caractéristiques complémentaires essentielles qui améliorent significativement la détection lorsqu'elle est combinée avec les données RGB.

Troisièmement, nous abordons les défis critiques de déploiement en développant un système temps réel et efficace sur le plan computationnel pour la détection automatisée de la région d'intérêt (ROI) et le filtrage des mouvements cliniquement non pertinents. Nous avons employé un réseau de détection basé sur des boîtes englobantes orientées qui localise précisément la région thoraco-abdominale, atteignant une précision moyenne (mAP) de 84% aux seuils IoU de 0,5 à 0,95. Cette approche orientée surpasse les méthodes traditionnelles alignées sur les axes en réduisant les fausses activations causées par l'équipement médical environnant et les artefacts environnementaux. De plus, nous proposons un détecteur de mouvements cliniquement non pertinents basé sur le flux optique et tenant compte des régions, qui atteint un score F_1 de 93% dans l'identification des segments vidéo où les symptômes de tirage sont difficiles à observer, garantissant que le système se concentre exclusivement sur les périodes diagnostiquement pertinentes.

Cette thèse présente une méthodologie complète pour un système de détection automatisée de la DRA. Le système proposé valide la faisabilité d'une surveillance respiratoire objective et continue, offrant le potentiel de réduire la charge de travail des cliniciens, d'améliorer la fiabilité diagnostique et de permettre la surveillance dans des contextes de soins de santé aux ressources limitées.

Mots-clés: Détection de la détresse respiratoire aiguë, évaluation médicale automatisée, rétractions thoraciques, surveillance du patient sans contact, apprentissage profond, imagerie RGB-D, réseaux de neurones convolutifs 3D, fusion de données multimodales

ASSESSMENT OF THE ACUTE RESPIRATORY DISTRESS USING A DEPTH CAMERA

Wajahat NAWAZ

ABSTRACT

Acute respiratory distress is an early phase of respiratory failure marked by severely impaired gas exchange, causing inadequate oxygenation of arterial blood and/or insufficient carbon dioxide removal, which can rapidly progress to respiratory failure if not treated promptly. This clinical emergency manifests through various observable signs that reflect the body's compensatory mechanisms attempting to maintain adequate oxygenation. Common clinical manifestations include tachypnea, nasal flaring, expiratory grunting, thoracoabdominal asynchrony and chest retractions. Among these clinical indicators, signs of chest retraction serve as highly specific and sensitive markers, as they directly indicate increased work of breathing. The presence of these signs constitutes a medical emergency necessitating prompt clinical intervention. The timely and accurate detection of these critical signs is paramount for initiating appropriate therapeutic interventions to prevent respiratory failure.

Current clinical practice relies primarily on visual assessment, a process where healthcare professionals physically observe patients at the bedside to score the severity of respiratory distress through identification of retraction signs. Visual assessment offers several advantages, including being non-invasive, providing immediate results, and requiring no specialized equipment. However, this manual, intermittent monitoring approach suffers from inter-observer variability and is resource-intensive, requiring continuous expert supervision. These limitations are particularly pronounced in resource-constrained settings and pandemic scenarios where clinical resources are strained.

This thesis presents an artificial intelligence-based contactless acute respiratory distress (ARD) detection system that mitigates the deficiencies of visual examination by automating the assessment process. The proposed system leverages RGB-D (color and depth) camera technology to capture visual and temporal information of the patient's chest wall in a non-invasive and continuous manner. The system further utilizes deep learning models to accurately localize chest wall regions and segment clinically meaningful temporal windows while effectively removing the motion artifacts. Advanced video analysis algorithms subsequently extract discriminative spatiotemporal features from the refined multi-modal data streams for automated respiratory distress identification.

This thesis makes three key contributions through interconnected studies. First, we evaluate various deep learning-based video analysis architectures for ARD detection in case of limited clinical data settings. Our evaluation reveals that real-world clinical datasets exhibit inherent spatial biases. To address this challenge, we propose a spatial-temporal selection framework. Systematic evaluation demonstrates that clinically relevant regions and appropriate temporal window length are critical for accurate and computationally efficient detection. Further analysis

reveals that models performing temporal downsampling alongside spatial feature extraction demonstrate superior performance compared to architectures that retain full temporal information. The proposed ARD system, leveraging the CSN-R101 model, attains an accuracy of an accuracy of 82%, precision of 80%, recall of 89%, and F_1 score of 84

Second, we investigate multi-modal data fusion for enhanced detection accuracy. We first establish that depth information alone is insufficient for robust ARD detection. Subsequently, we demonstrate that late feature fusion of RGB and depth modalities substantially outperforms single-modality approaches, achieving 85.2% accuracy, 86.7% precision, 85.2% recall, and 85.8% F_1 score, significantly improving upon the RGB-only (82.2% accuracy, 87.2% precision, 77.7% recall, 82.1% F_1 score). These findings demonstrate that while depth alone is inadequate, it provides essential complementary features that significantly improve detection when combined with RGB data

Third, we address critical deployment challenges by developing a real-time, computationally efficient system for automated region-of-interest (ROI) detection and filtering of clinically irrelevant movements. We employed an oriented bounding box-based detection network that precisely localizes the thoracoabdominal region, achieving an 84% mean Average Precision (mAP) at IoU thresholds 0.5 to 0.95. This oriented approach outperforms traditional axis-aligned methods by reducing false activations caused by surrounding medical equipment and environmental artifacts. Additionally, we propose an optical flow-based, region-aware clinically irrelevant movement detector that attains a 93% F_1 score in identifying video segments where retraction symptoms are difficult to observable, ensuring the system focuses exclusively on diagnostically relevant periods.

This thesis presents a comprehensive methodology for automated acute respiratory ARD detection system. The proposed system validates the feasibility of objective, continuous respiratory monitoring with the potential to reduce clinician workload, improve diagnostic reliability, and enable monitoring in resource-constrained healthcare settings.

Keywords: Acute Respiratory Distress Detection, Automated Medical Assessment, Chest Retractions, Contactless Patient Monitoring, Deep Learning, RGB-D Imaging, 3D-Convolutional Neural Networks, Multi-modal Data Fusion

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

Page

LIST OF ABBREVIATIONS

AR	Auto-Regressive
ARD	Acute Respiratory Distress
Adam	Adaptive Moment Estimation
CAM	Class Activation Map
CNN	Convolutional Neural Networks
CSN	Channel-Separated Networks
FP	False Positive
FN	False Negative
FCAT-F	Feature Concatenation with Frozen base models
FTA	Face Thoracic-abdominal
FPS	Frames Per Second
GRU	Gated Recurrent Unit
HM	Hermite Magnification
ICA	Independent Component Analysis
ICU	Intensive Care Units
I:E	Inspiratory/Expiratory Ratio
LSTM	Long Short-Term Memory
mAP	mean Average Precision
MDR	Microwave Doppler Radar

MV	Minute Ventilation
NICU	Neonatal Intensive Care Units
PICU	Pediatric Intensive Care Units
R	Correlation Coefficient
REB	Research Ethics Board
RGB	Red, Green, Blue (color channels)
RGB-D	Red, Green, Blu &-Depth
RNN	Recurrent Neural Networks
ROI	Region of Interest
RR	Respiratory Rate
SGD	Stochastic Gradient Descent
S3D	Separable 3D Convolution
SNR	Signal-to-Noise Ratio
SWIN-VIT	Shifted Window Vision Transformer
TAA	Thoracoabdominal Asynchrony
TN	True Negative
TP	True Positive
Vt	Tidal Volume
X3D	Efficient 3D-CNN Architecture
OBB	Oriented Bounding Box

LIST OF SYMBOLS AND UNITS OF MEASUREMENTS

τ_m	Motion Magnitude Threshold
α	Learning Rate Parameter
$\mathcal{M}_{\text{bin}}(x, y)$	Binary Mask
β	Momentum or Regularization Parameter
λ	Weighting Factor or Regularization Coefficient
σ	Standard Deviation
$\mathcal{P}_{\text{moving}}$	Percentage of Moving Pixels
τ_p	Percentage Threshold
μ	Mean Value
ϵ	Error or Threshold Value
Δ	Change or Difference
∞	Infinity
$\approx$	Approximately Equal To

INTRODUCTION

Pediatric Intensive Care Units (PICUs) provide specialized critical care for patients typically ranging from newborns to 18 years of age. These vulnerable patients require continuous physiological monitoring and immediate medical intervention to prevent further health complications. Medical instruments play a pivotal role in this environment by monitoring patient physiological parameters real-time, assisting clinicians in accurate diagnosis, enabling early detection of clinical deterioration, guiding therapeutic decision-making, and evaluating treatment efficacy. Clinicians continuously monitor multiple physiological parameters to assess patient status, including cardiovascular, respiratory, and neurological indicators. Cardiovascular parameters include heart rate, blood pressure, and cardiac rhythm patterns, which provide information about circulatory function. Respiratory parameters include respiratory rate, oxygen saturation, and carbon dioxide levels, reflecting pulmonary gas exchange efficiency. Neurological parameters include consciousness levels and, in specific cases, intracranial pressure to assess neurological status. While comprehensive monitoring of all physiological parameters is essential, respiratory function assessment is particularly critical, as respiratory failure represents a leading cause of PICU admission and mortality (Donoso, Arriagada, Contreras, Ulloa & Neumann, 2016).

Respiratory failure is the leading cause of cardiac arrest in children. Pediatric patients with acute respiratory distress (ARD) can deteriorate rapidly, progressing from breathing difficulties to respiratory failure and ultimately cardiac arrest without timely intervention (Kumar & Vishnu, 1996; Behrman, Kliegman, Jenson *et al.*, 2000). Therefore, early recognition of ARD is critical for timely clinical intervention and prevention of life-threatening complications. ARD can be detected through complementary modalities: imaging (chest X-ray and computed tomography), physiological parameters (partial pressure of oxygen and carbon dioxide), and visual examination. Each modality offers unique diagnostic value. Chest X-ray (CXR) serves as the primary imaging tool, identifying bilateral infiltrates, pulmonary edema, and consolidation patterns indicative of ARD or ARD syndrome. Computed tomography (CT) scans provide superior spatial resolution

for detailed visualization of lung parenchyma. CT enables assessment of ground-glass opacities and interstitial changes, particularly when CXR findings are inconclusive. Arterial blood gas analysis quantifies gas exchange by measuring partial pressure of oxygen (PaO2) and carbon dioxide (PaCO2), revealing hypoxemia or hypercapnia. Pulse oximetry enables continuous, non-invasive SpO2 monitoring. Capnography assesses end-tidal CO2 for ventilation adequacy (Hon, Leung, Oberender & Leung, 2021; Payán, Pedroza-Granados & Domínguez-Cherit, 2007).

Clinical observation detects immediate visual signs of respiratory distress, including restlessness or confusion (signaling hypoxemia or hypercapnia), cyanosis of lips and nail beds (indicating severe oxygen deprivation), nasal flaring and chest retractions (showing increased work of breathing), use of accessory muscles (reflecting respiratory effort), and abnormal lung sounds (suggesting airway obstruction) (Diamond, Peniston, Sanghavi & Mahapatra, 2025; Yehya & Thomas, 2016; Hammer, 2013; Vargas-Acevedo *et al.*, 2024; Blonski, 2025). Despite the diagnostic value of imaging and laboratory studies, visual examination remains the preferred initial assessment method in pediatric populations. It is non-invasive, requires no specialized equipment, enables rapid bedside evaluation, and avoids radiation exposure and patient discomfort associated with radiological procedures. This makes visual examination especially suitable for frequent monitoring of vulnerable pediatric patients who may be unable to tolerate invasive diagnostic interventions.

Among clinical visual signs, chest retractions represent the most critical indicator of ARD, enabling rapid, non-invasive early detection of respiratory compromise. Chest retractions are defined as the abnormal inward movement of the soft tissues of the chest wall during inspiration, resulting from excessive negative intrathoracic pressure. These retractions can occur in intercostal (between ribs), subcostal (below ribs), or suprasternal (above the sternum) regions. Retraction signs indicate that the respiratory system is experiencing significant airway obstruction or reduced lung compliance, forcing the patient to generate excessive inspiratory effort to maintain

adequate ventilation. However, prolonged use of accessory muscles leads to respiratory muscle fatigue, resulting in impaired ventilatory function and eventual respiratory failure. The location, severity, and duration of retractions provide critical diagnostic information regarding the degree of airway obstruction and the risk of impending respiratory failure (Ernstmeyer & Christman, 2021; Edwards, Kotecha & Kotecha, 2013a; Warren & Anderson, 2010).

Despite significant advancements in medical technology and monitoring capabilities, critical visual indicators of respiratory distress in pediatric intensive care settings are still assessed through visual examination. Healthcare professionals must identify these indicators through direct visual assessment at the bedside during routine patient rounds, relying on their clinical expertise and observational skills to detect subtle changes in patient condition (Vitazkova *et al.*, 2024). This traditional approach, while well-established in clinical practice, has been the standard of care for decades due to its immediacy and non-invasive nature. However, reliance on visual assessment introduces three major limitations that can compromise patient safety and clinical outcomes. First, visual assessment provides only intermittent rather than continuous monitoring, creating temporal gaps during which critical changes in patient status may go undetected. Between scheduled patient rounds or nursing assessments, rapid deterioration can occur without immediate recognition, potentially delaying life-saving interventions. Second, substantial inter-observer variability exists in severity grading of retraction signs, as different clinicians may interpret the same clinical presentation differently based on their experience level, training background, and subjective judgment. Additionally, clinician fatigue and cognitive overload compromise diagnostic accuracy, which can lead to inconsistent documentation, delayed escalation of care, and potential miscommunication among healthcare team members. Third, effective continuous visual monitoring requires constant expert presence at the bedside, which is often impractical given staffing shortages and high patient-to-nurse ratios.

In resource-constrained settings and during pandemics, these limitations have far more serious consequences. Many rural hospitals, community health centers, and facilities in developing countries face severe shortages of trained pediatric specialists and intensive care personnel (Ahmed *et al.*, 2025). High patient-to-clinician ratios force them to divide their attention among multiple critically ill patients simultaneously, making continuous bedside monitoring practically impossible and increasing the risk of delayed recognition of respiratory deterioration (Milesky, Sharma, Rosen, Zhang & Bass, 2025). Telemedicine consultations face similar constraints, as remote clinicians cannot continuously observe patients or objectively quantify visual respiratory signs such as chest retractions (Kapus, Rárosi, Novák, Peták & Tolnai, 2025). Even well-resourced facilities experience workforce strain during pandemic situations or mass-casualty events, compromising their monitoring capacity (Douglas, Mehta & Mansoori, 2024). These challenges underscore the urgent need for an automated, objective, and continuous monitoring system capable of detecting acute respiratory distress by analyzing chest deformation without requiring constant human observation. Such automation could provide real-time alerts, reduce clinician workload, eliminate inter-observer variability, and enable remote monitoring capabilities to address critical needs in both resource-limited and high-acuity care environments (Health Recovery Solutions, 2025)

Recent advances in computer vision have transformed the medical field by enabling automated analysis of visual information that previously required expert interpretation (Vesal, Ravikumar, Davari, Ellmann & Maier, 2018; Cireşan, Giusti, Gambardella & Schmidhuber, 2013; Lanjewar, Panchbhai & Charanarur, 2023; Sultani *et al.*, 2022; Acharya & Basu, 2020). Computer vision systems can now identify subtle patterns in medical images that indicate disease, often matching or exceeding the diagnostic accuracy of trained specialists. For example, these systems can detect pneumonia in chest X-rays (Gabruseva, Poplavskiy & Kalinin, 2020; Mazurowski, Buda, Saha & Bashir, 2019), identify cancerous cells in tissue samples (Wang *et al.*, 2022; Ahmad *et al.*, 2022), and recognize early signs of diabetic retinopathy in retinal photographs (Saini & Susan,

2022). The key advantage of these systems is their ability to provide consistent, objective assessments while operating continuously without fatigue, making them particularly valuable for high-volume screening and continuous monitoring applications. Building upon these successes in medical image analysis, researchers have begun applying computer vision techniques to automate respiratory monitoring tasks in a contactless and non-invasive manner. These contactless vision-based methods offer several key advantages for pediatric patients. First, they provide completely non-invasive, contactless operation that eliminates physical sensors and attachments, thereby enhancing patient comfort and reducing risks of skin irritation, pressure injuries, and device-related infections. This is particularly beneficial for pediatric patients who may exhibit resistance to traditional monitoring equipment or experience heightened anxiety from physical contact with medical devices. Second, these approaches enable continuous monitoring in unconstrained clinical environments using affordable consumer-grade cameras, effectively addressing practical challenges including limited patient mobility, equipment costs, and resource constraints commonly encountered in pediatric intensive care settings. Prior studies utilizing RGB, depth, and RGB-D cameras have demonstrated the feasibility of estimating respiratory rate through chest wall movement tracking (Tarassenko *et al.*, 2014), detecting thoracoabdominal asynchrony (Huang *et al.*, 2024), and measuring tidal volume (Rehouma, Noumeir, Bouachir, Jouvet & Essouri, 2018).

Inspired by the success of contactless monitoring methods in respiratory assessment, this thesis presents an automated ARD detection system based on RGB-D camera technology and advanced computer vision algorithms. The proposed system leverages the complementary strengths of RGB and depth modalities, integrated with advanced computer vision techniques, to enable continuous, real-time ARD monitoring, addressing critical limitations in current manual assessment methods.

Research Objectives

The main goal of this research is to design an automated framework for identifying acute respiratory distress (ARD) in pediatric intensive care units (PICUs).

To achieve this, we define three short term objectives:

1. **Investigate video understanding techniques for ARD detection:** The research goal is to benchmark video analysis architectures for automated ARD detection from video data.
2. **Develop multi modal fusion strategies:** Our second objective is to leverage both RGB and depth modalities to improve the ARD detection system accuracy.
3. **Design automated spatiotemporal preprocessing methods:** Develop automated methods to identify the thoracic abdominal (region of interest) and detect clinical irrelevant movements.

By achieving these objectives, this research establishes a foundation for automated acute respiratory distress monitoring in clinical setting. This approach reduces the burden on healthcare staff while enabling continuous patient surveillance.

Research Contributions

This work contributes through three interconnected studies, each building upon the previous investigation to progressively advance automated respiratory distress detection capabilities.

Contribution 1: Automated Detection of Acute Respiratory Distress Using Temporal Visual Information

This work establishes the feasibility of automated ARD detection using RGB video analysis and addresses fundamental challenges in applying deep learning to PICU environments. First, it formalizes ARD detection as a video classification problem and designs a data acquisition

protocol suitable for clinical settings. Second, it systematically investigates preprocessing strategies to address the limited medical data challenge. This study shows that models trained on thoracic-abdominal region data achieve significantly better performance (accuracy: 72.5% to 81.2%) compared to weakly processed data. This improvement is attributed to the models learning clinically relevant features rather than irrelevant background information. Third, it identifies optimal temporal parameters through systematic experiments. We selected 6.4-second video clips at 10 FPS (frame per second) for ARD detection. Pediatric respiratory rates vary widely: adolescents breathe 12-20 times per minute, while newborns breathe 30-60 times per minute. The 6.4-second temporal window ensures acquisition of at least one complete respiratory cycle across the entire pediatric age range while optimally balancing temporal coverage and computational efficiency.

Fourth, it establishes baseline performance by comparing multiple video understanding architectures, including 2D-CNNs+LSTM and various 3D-CNNs (R(2+1)D, S3D, SlowFast, X3D, CSN, SWIM-VIT, CSN-R101). Among these architectures, CSN-R101 achieves optimal performance with 81.9% accuracy, 79.8% precision, 89.1% recall, and 84.0% F_1 score. Class activation map analysis validates that trained models focus on clinically relevant thoracic-abdominal regions for decision making, demonstrating the viability of automated ARD detection for continuous monitoring in resource-constrained settings.

Contribution 2:Acute Respiratory Distress Identification via Multi-Modality Using Deep Learning

This work addresses the limitations of single-modality approaches by leveraging multi-modality (RGB-D) data. This study demonstrates that integrating depth information with RGB overcomes the limitations of RGB modality, which is sensitive to skin tone and lighting conditions. However, the depth modality alone cannot be used for ARD detection due to its inability to capture visual appearance features such as texture patterns that are critical indicators of respiratory distress.

Second, it develops depth modality preprocessing techniques that removes outliers and scales depth values for optimal integration with RGB data. Third, it investigates multiple fusion strategies including early fusion, late feature fusion, score averaging, and feature concatenation with frozen base models (FCAT-F) to effectively combine both modalities. Experimental results show that FCAT-F achieves optimal performance: 85% accuracy, 87% precision, 85% recall, and 85.8% F_1 score, representing significant improvement over single-modality approaches and demonstrating more robust performance in clinical settings.

Contribution 3: Improving Acute Respiratory Distress Detection in Pediatric ICUs Using Automated Spatiotemporal Region Extraction

This work advances automated acute respiratory distress detection system by developing a fully automated spatiotemporal preprocessing pipeline that eliminates manual intervention while addressing fundamental biases inherent in clinical video data. First, we develop an oriented bounding box (OBB) based ROI detection system for automated chest region localization in PICU settings. Unlike axis-aligned boxes constrained to horizontal-vertical edges, OBBs incorporate rotation parameters to align with the infant's body axis regardless of positioning. This method achieves 84% mAP@50-95 (mean average precision across IoU thresholds from 0.5 to 0.95) and offers two key advantages: (1) it minimizes irrelevant background inclusion, and (2) it removes spatial biases introduced by medical equipment artifacts more effectively. Second, we propose a region-aware motion detection network that automatically identifies and excludes video segments containing excessive clinically irrelevant movements, such as patient repositioning, hand movements, and crying episodes. This temporal filtering component achieves 93% F_1 score in distinguishing between respiratory motion and non-clinical movements, preventing models from learning spurious correlations with irrelevant patient activity. Third, we establish both quantitative and qualitative evidence of the spatiotemporal region extraction's impact on ARD detection performance through extensive ablation studies. Our results demonstrate

that clinical region isolation and removal of irrelevant movement clips are both necessary for reliable diagnosis, with precision improving from 76% to 80% and specificity from 72% to 78% compared to axis-aligned and poorly processed approaches. Class activation map analysis provides visual confirmation that our preprocessing robustly redirects model attention from confounding artifacts to clinically relevant respiratory regions. Together, these contributions enable continuous automated patient monitoring in intensive care units under real-world conditions without requiring manual intervention.

Overall Impact and Thesis Structure

This thesis establishes the first automated system for continuous acute respiratory distress detection based on chest retraction signs in pediatric intensive care, addressing a critical clinical gap where detection was previously dependent on subjective, intermittent visual examination.

Clinical Impact: The system enables continuous, objective respiratory assessment without requiring constant clinician presence, directly addressing workforce limitations in resource-constrained environments. Rural hospitals, facilities in developing countries, and centers experiencing specialist shortages face significant challenges in providing continuous expert monitoring. By automating visual assessment procedures, this system reduces risks of delayed detection, eliminates inter-observer variability in severity assessment, and enables remote monitoring capabilities previously impossible with manual visual examination alone.

Technical Contributions: Deep learning models learn from training data patterns, but PICU videos contain substantial non-clinical information such as medical equipment, monitoring devices, and background artifacts that can mislead models. This research demonstrates that strategic preprocessing is essential to guide models toward clinically relevant features and prevent learning from non-clinical information. Two technical innovations address this challenge: (1)

automated spatiotemporal region of interest extraction and (2) incorporating multi-modality information for ARD detection.

Broader Implications: Beyond respiratory monitoring, this work establishes methodological foundations with broader clinical applications. The automated preprocessing pipeline addresses a fundamental challenge in medical deep learning: extracting clinically relevant information from complex, noisy video data. This framework is essential for robust model performance and successful clinical deployment. The region-aware motion detection system developed for filtering non-respiratory movements has additional applications. It can identify false alarms in PICU monitoring systems caused by patient repositioning, crying, or other non-clinical activities.

Thesis Organization: The remainder of this thesis is organized as follows: Chapter 1 reviews related work in automatic ARD detection, respiratory monitoring and advance video analysis algorithms. Chapters 2, 3, and 4 present the three studies in detail. Finally, the thesis discusses clinical implications, limitations, and future directions.

CHAPTER 1

LITERATURE REVIEW

This chapter presents a comprehensive review of existing literature on acute respiratory distress (ARD) detection and automated respiratory assessment methodologies. The review is structured in three parts: First, it examines traditional ARD diagnostic approaches leveraging medical imaging modalities (chest X-ray and CT scans) and physiological signal-based monitoring systems and presents the limitations of these methods. Second, it explores recent advances in contactless respiratory monitoring for pediatric intensive care units (PICUs), emphasizing RGB camera-based methods, depth sensor-based approaches, and multimodal fusion techniques, and identifies critical research gaps. Third, it provides a comprehensive literature review of advanced video analysis algorithms, including 2D CNNs with recurrent networks, 3D CNNs, and vision transformers.

1.1 Acute Respiratory Distress Detection

Traditional diagnostic approaches for acute respiratory distress have long relied on established clinical modalities such as chest radiography (X-ray) and computed tomography, as well as clinical data including oxygen saturation levels in blood and respiratory rate. These modalities provide objective measurements of respiratory function and anatomical changes (Hart & Lee, 2019), which are further interpreted by experts for diagnosis. Recent research has automated the interpretation process by analyzing individual and multi-modality information using advance computer vision techniques, aiming to accelerate clinical diagnosis.

Pardasani et al. (Pardasani, Chaudhuri, Awasthi & Goel, 2020) proposed an automatic system to detect acute respiratory distress using clinical data including respiratory rate (RR) and peripheral capillary oxygen saturation (SpO_2). They employed machine learning model to quantify respiratory distress into a real-time by analyzing clinical data. The study evaluated several methodologies on 912 records from the MIMIC-III Clinical and Waveform Database, including logistic regression, decision tree, support vector machine, multi-layer perceptron,

convolutional neural network (CNN), and long short-term memory networks. All models achieved an area under the curve for the receiver operating characteristic (AUC) either close to or above 90% for the respiratory distress detection task. The CNN model demonstrated marginally better performance across all metrics, achieving sensitivity and specificity of 86% and 85%, respectively. Their system address the problem of alarm fatigue by making prediction using multiple clinician parameters.

Reamaroon et al. (Reamaroon *et al.*, 2021) proposed an automated acute respiratory distress syndrome (ARDS) detection system by analyzing chest X-rays (CXRs). Their method extracted the lung region from chest X-ray images using total variation-based active contour methods for ARDS detection. An AdaBoost machine learning model was trained using four feature sets: directionality measures, first-order histogram statistics, gray-level co-occurrence matrix features, and deep learning features from pre-trained neural networks. Experimental results demonstrated that the multi-feature model achieved the best overall performance, yielding an accuracy of 83% and an Area under the Curve (AUC) of 79% compared to individual feature models.

Whereas the previous study employed chest X-ray imaging, Grooby et al. (Grooby *et al.*, 2022) investigated respiratory distress detection in term newborn babies using an acoustic modality: digital stethoscope sound recordings captured within one minute post-delivery. Their method preprocessed chest sounds to isolate high-quality heart and lung sounds, then extracted power and vital sign features. These features were classified using the hybrid sampling/boosting algorithm (RUSBoost) algorithm (Seiffert, Khoshgoftaar, Van Hulse & Napolitano, 2009). When evaluated on a dataset of 51 term newborns, the approach achieved 85.0% specificity, 66.7% sensitivity, and 81.8% accuracy. These results demonstrate the feasibility of using digital stethoscope recordings for early, non-invasive detection of respiratory distress immediately following delivery.

Sjoding et al. (Sjoding *et al.*, 2021) leveraged deep learning and transfer learning techniques to automate ARDS detection from chest X-ray images. Their training approach consisted of two stages: pre-training on over 595,000 annotated chest radiographs to learn general radiographic features, followed by fine-tuning on ARDS-specific expert-labeled data. This approach achieved

robust performance with an Area Under the Receiver Operating Characteristic curve (AUROC) of approximately 92% on internal validation and approximately 88% on external validation, outperforming clinicians while demonstrating strong calibration. This work validated the capacity of CNN-based models to directly interpret raw chest radiographs for reliable ARDS detection, representing a notable advance over prior approaches that relied on predefined features.

Moving beyond single-modality approaches, Pai et al. (Pai *et al.*, 2022) developed a multimodal AI framework for ARDS diagnosis by integrating clinical data and chest radiographs through ensemble learning. Their method employed CNN models to predict ARDS from chest X-ray images and machine learning algorithms (XGBoost, Random Forest, and Logistic Regression) from clinical data. The system aggregated the predictions from both the modality using weighted ensemble aggregation technique for final diagnosis. Experimental results demonstrated that ensemble averaging achieved an AUC of 0.920±0.02, sensitivity of 0.846±0.02, and specificity of 0.863±0.02. By incorporating both chest X-rays and clinical data, the system enabled more accurate and explainable ARDS identification aligned with the Berlin definition.

Fonck et al. (Fonck, Fritsch, Nottenkämper & Stollenwerk, 2023) implemented a ResNet-50 deep learning model with transfer learning to detect ARDS from chest CXRs images, aiming to accelerate clinical diagnosis. The approach leveraged the ResNet-50 architecture pretrained on large image datasets and fine-tuned specifically on ARDS CXRs dataset. The model achieved an AUC of 92.6%, sensitivity of 87%, and specificity of 97%, demonstrating effective ARDS classification performance and supporting faster diagnostic processes in critical care settings.

Farzaneh et al. (Farzaneh, Ansari, Lee, Ward & Sjoding, 2023) developed a collaborative system to assist clinicians in ARDS detection from chest X-rays. Rather than focusing solely on model performance, they investigated four AI-physician collaboration strategies: AI-first review with deferral of uncertain cases to physicians, physician-first review and defer to AI, and solo approaches by either AI or clinician alone. Their AI model initially outperformed solo clinicians in standalone evaluation. The AI-first collaborative strategy achieved the highest accuracy of 87%, surpassing both physician-alone (81%) and AI-alone (85%) approaches. This

strategy achieved optimal performance by allowing the AI to autonomously diagnose 79% of straightforward cases while reserving physician review for the 21% of cases with equivocal features.

Zhang et al. (Zhang & Pang, 2023) developed predictive machine learning models to identify acute pancreatitis (AP) patients at high risk of developing ARDS using clinical and laboratory biomarkers. Their retrospective study included 460 AP patients admitted between January 2017 and August 2022, with 83 (18.04%) subsequently developing ARDS. Their methods use eight clinical data features: partial pressure of oxygen, C-reactive protein, procalcitonin, lactic acid, calcium ion concentration, neutrophil-to-lymphocyte ratio, white blood cell count, and amylase levels. Four predictive models were evaluated: support vector machine, ensembles of decision trees, bayesian classifier, and a nomogram. The models were trained using 5-fold cross-validation and subsequently evaluated on an independent test dataset. The Bayesian classifier achieved the highest discrimination performance with an AUC of 89%, surpassing support vector machine (87%), ensemble decision trees (81%), and the nomogram (87%). However, ensemble decision trees demonstrated superior classification metrics with 89% accuracy, 80% precision, and 62% F_1 score, along with the lowest false discovery rate (20%) and highest negative predictive value (90%), suggesting strong potential for clinical deployment.

Yahyatabar et al. (Yahyatabar, Jouvet & Cheriet, 2023) advanced ARDS detection by proposing Dense-Ynet, an interpretable deep learning architecture that jointly performs classification and segmentation of chest X-rays within a clinically aligned framework. The model partitions each lung into quadrants (upper and lower regions) to evaluate bilateral diffuse infiltrates, mirroring the diagnostic criteria used in clinical ARDS assessment, and achieved an AUROC of 95%. In contrast to black-box methods, this approach provides transparent decision-making by replicating the diagnostic workflow employed by clinicians, thereby addressing a critical barrier to clinical adoption of AI diagnostic systems. The framework demonstrated robust performance in both lung region segmentation and quadrant-based classification, maintaining diagnostic accuracy even in pathological images where lung boundaries were poorly defined.

Table 1.1 Summary of ARDS Detection Methods and Performance

Study	**Modality**	**Method**
Pardasani *et al.* (2020)	Clinical data	CNN, LSTM, Bayesian inference
Reamaroon *et al.* (2021)	Chest X-ray	AdaBoost, Deep model features
Grooby *et al.* (2022)	Digital Stethoscope Sound	RUSBoost
Sjoding *et al.* (2021)	Chest X-ray	2DCNN
Pai *et al.* (2022)	Chest X-ray, Clinical data	Ensemble (2DCNN + ML algorithms)
Fonck *et al.* (2023)	Chest X-ray	2DCNN
Farzaneh *et al.* (2023)	Chest X-ray	2DCNN Physician Collaboration
Zhang & Pang (2023)	Clinical data	SVM, EDTS, BC, Nomogram
Yahyatabar *et al.* (2023)	Chest X-ray	Dense-Ynet
Zhou *et al.* (2024)	CT Scan	UNETR

CNN: Convolutional Neural Network; LSTM: Long Short-Term Memory; RUSBoost: Random Under-Sampling Boosting; 2DCNN: Two-Dimensional Convolutional Neural Network; ML: Machine Learning; SVM: Support Vector Machine; EDTS: Ensembles of Decision Trees; BC: Bayesian Classifier; UNETR: UNEt TRansformers

Zhou et al. (Zhou *et al.*, 2024) developed a multimodal deep learning framework for ARDS detection that integrates quantitative CT imaging analysis with clinical metadata. The study utilized data from 928 ICU patients across three hospitals to train and validate a UNet Transformer (UNETR) architecture for automated lung segmentation and ARDS prediction. The approach utilized a UNet Transformer (UNETR) architecture, achieving a lung segmentation

Dice coefficient of 0.734±0.137 and ARDS prediction AUROCs ranging from 0.865 to 0.916 across internal, external, and prospective validation cohorts. The model combined quantified lung lesion features (including ground-glass opacity, consolidation, pulmonary fibrosis, and pleural effusion) extracted from segmented CT scans with clinical metadata to predict ARDS development. Model interpretability was enhanced through Shapley explanation plots, revealing the relative contribution of imaging and clinical features to prediction outcomes. This framework demonstrated superior performance compared to DenseNet-based image classification methods and exhibited robust generalization across diverse validation datasets, facilitating early identification of high-risk patients for timely clinical intervention.

Research GAP

Table 1.1 summarizes the existing ARDS detection methods across different modalities and their corresponding approaches. While existing ARD detection methods demonstrate strong performance, they face significant limitations in real-time monitoring. Current approaches predominantly rely on chest X-rays imaging (Reamaroon *et al.*, 2021; Sjoding *et al.*, 2021; Pai *et al.*, 2022; Fonck *et al.*, 2023; Farzaneh *et al.*, 2023; Yahyatabar *et al.*, 2023; Zhou *et al.*, 2024), which require specialized equipment, trained personnel which is a time-consuming processes. Similarly, methods utilizing clinical metadata depend on invasive blood sampling and laboratory analysis (Zhang & Pang, 2023), introducing diagnostic delays. Although a few studies have explored oxygen saturation level, respiratory rate and sound parameters captured by specialized equipment for real-time ARDS detection (Pardasani *et al.*, 2020; Grooby *et al.*, 2022), widespread adoption remains limited. These dependencies on complex imaging modalities and laboratory-based data prevent timely ARD detection, particularly in emergency and resource-limited settings where rapid intervention is critical. This highlights the need for non-invasive, real-time detection methods that enable immediate assessment without specialized equipment or extensive processing time.

1.2 Contactless Monitoring Methods

The limitations inherent in contact-based methods motivated researchers to develop contactless alternatives. Early contactless approaches leveraged standard RGB and RGB-D cameras due to several practical advantages: widespread clinical availability, non-invasive operation, and capacity to capture comprehensive visual information regarding patient position and movement patterns. These methods employed signal processing and computer vision techniques to extract respiratory parameters from video data.

1.2.1 RGB Camera-Based Contactless Monitoring

Initial RGB camera based approaches leveraged temporal differencing for respiratory motion detection. Bai et al. (Bai, Li & Chen, 2010) developed a contactless respiratory rate estimation system employing two RGB cameras to capture the patient's torso from two viewpoints. Their method computed absolute frame-to-frame differences to identify motion regions, with respiratory rate derived by analyzing periodic motion intensity patterns within the thoracic region. The system detected abnormal respiratory rates and generated alarms when respiration stopped for more than 10 seconds. Despite computational simplicity enabling real-time processing, temporal differencing exhibited significant limitations, including high sensitivity to non-respiratory movements such as patient repositioning, hand gestures, shadows, and lighting variations, which complicated isolation of genuine respiratory motion from confounding temporal patterns.

Advancing beyond simple temporal differencing, Bartula et al. (Bartula, Tigges & Muehlsteff, 2013) proposed a computationally efficient system employing vertical profile projection combined with cross-correlation to detect chest wall deformations. This approach provided directional information that distinguished inhalation from exhalation, enabling real-time breathing waveform reconstruction. A key innovation was the integration of large motion detection capabilities that automatically identified and excluded time segments containing non-respiratory movements, such as patient repositioning or arm gestures. The system employed a breath-to-breath classification scheme using decision trees to label individual respiratory cycles as "good" or "bad," with only

validated breaths contributing to rate calculations. Validation on a mechanical phantom (9,800 breathing cycles) and five healthy volunteers demonstrated strong performance with correlation coefficients of R=0.97 and R=0.98, respectively, achieving 95% precision after post-processing. However, the study was limited to controlled laboratory settings with healthy subjects and did not address detection of respiratory distress indicators, such as chest retractions, or automated region of interest localization, which remained unmet clinical needs.

Tarassenko et al. (Tarassenko *et al.*, 2014) developed a video-based vital sign monitoring system employing auto-regressive (AR) signal processing to analyze subtle pixel intensity variations in facial regions captured via standard RGB cameras. AR models enhanced signal quality and noise suppression relative to frequency-domain approaches such as fast Fourier transform, with temporal modeling of physiological signal dynamics providing improved handling of signal irregularities and motion artifacts. The system achieved respiratory rate estimation with a mean absolute error (MAE) of 2.32 breaths per minute and root mean square error (RMSE) of 3.39 breaths per minute when validated on 10 adult volunteers under diverse ambient lighting conditions, confirming reliable vital sign extraction without specialized illumination. However, clinical applicability for respiratory distress assessment faced significant constraints. The approach focused on facial photoplethysmographic effects rather than direct thoracic observation, and consequently, spatial respiratory effort patterns including chest retractions and regional breathing abnormalities critical for distress assessment remained undetectable.

Al-Naji et al. (Al-Naji & Chahl, 2016) proposed a remote respiratory monitoring approach that employed motion magnification techniques to enhance subtle chest wall movements. Specifically, the method utilized Eulerian Video Magnification (EVM) (Liu, Lu, Luo, Zhang & Chen, 2014), which amplifies minor motion changes within spatial and temporal frequency bands associated with breathing. This process made otherwise imperceptible respiratory motions visible for algorithmic interpretation. The magnified recordings were subsequently processed to select regions of interest, followed by temporal signal analysis to derive respiratory rate. Tests conducted on healthy participants demonstrated that motion magnification improved the signal-to-noise ratio of respiratory patterns, particularly under shallow breathing or partial

occlusion conditions. However, the computational complexity of the EVM framework limited its real-time applicability, and the uniform amplification of motions within chosen frequency bands, including camera noise and unrelated body movements introduced potential artifacts and reduced measurement reliability.

Mateu et al. (Mateu-Mateus, Guede-Fernandez, Garcia-Gonzalez, Ramos-Castro & Fernández-Chimeno, 2020) employed dense optical flow algorithms for respiratory rhythm extraction, representing a more sophisticated computer vision approach to motion analysis. Dense optical flow computes motion vectors for every pixel in the image between consecutive frames, providing comprehensive spatial information about chest wall displacement patterns rather than aggregate motion estimates. Their system, positioned at the patient's lateral side, analyzed the vertical component of these dense motion vectors in thoracic and abdominal regions to extract respiratory signals. This dense representation enabled capture chest expansion and lateral rib cage movements with greater spatial granularity than previous temporal differencing or sparse feature tracking methods. Respiratory rate was extracted by identifying peaks in the motion signal corresponding to breathing cycles. The approach offered practical advantages, including easier camera placement in clinical settings and reduced facial occlusion. However, dense optical flow computation more computationally intensive than simpler methods, remained sensitive to non-respiratory patient movements, and required careful positioning to maintain the thoracic region within the field of view. The study focused exclusively on respiratory rate estimation in healthy subjects and did not leverage the rich spatial motion information to detect other clinical indicators of respiratory distress.

Massaroni et al. (Massaroni, Lo Presti, Formica, Silvestri & Schena, 2019) proposed a non-contact respiratory monitoring system based on a conventional RGB webcam, where temporal variations of pixel intensity over a region interest are used to derive a respiratory signal and estimate breath-by-breath respiratory rate. The region of interest (ROI) is defined in the first frame by manually selecting a point at the jugular notch, around which a rectangular area is automatically constructed to cover the upper thoracoabdominal wall; within this ROI, RGB

intensities are averaged along horizontal lines, and only the lines with the highest temporal variability are retained to build the video-derived respiratory waveform.

Brieva et al. (Brieva, Ponce & Moya-Albor, 2020) proposed a contactless respiratory rate estimation method combining Hermite Magnification (HM) with convolutional neural networks (CNNs), representing an early integration of deep learning with motion enhancement techniques for respiratory monitoring. Their system utilized a single RGB camera to capture video of the patient's torso, with the magnified motion subsequently analyzed by a CNN architecture trained to classify individual frames as corresponding to either inhalation or exhalation phases. This frame-level classification approach enabled reconstruction of the breathing cycle and extraction of respiratory rate through temporal analysis of the classified sequence. The integration of CNNs marked a departure from traditional signal processing pipelines, leveraging learned feature representations rather than hand-crafted motion descriptors. Validation on ten healthy subjects demonstrated promising performance with an average error of 3.28 ± 3.33% for respiratory rate estimation. However, the system exhibited significant motion sensitivity, producing inaccurate readings when patients made small non-respiratory movements. Furthermore, the study was limited to controlled settings with healthy volunteers and relatively small sample size, and did not address detection of pathological breathing patterns or clinical indicators of respiratory distress beyond rate estimation. The reliance on frame level classification also required temporally consistent video quality, as individual frame misclassifications could propagate errors through the respiratory cycle reconstruction process.

Romano et al. (Romano, Schena, Silvestri & Massaroni, 2021) proposed a low-cost, contactless respiratory monitoring system using RGB images from a consumer-grade camera. The system employed automated chest tracking and two breathing signal extraction methods: optical flow (FO) for tracking chest wall displacement and RGB intensity analysis for monitoring pixel brightness variations. Validation with 24 healthy volunteers evaluated performance across varying camera distances (0.5-2 meters) and postures (sitting, standing, supine). The optical flow method consistently outperformed RGB intensity analysis across all tested conditions, the best performance achieved when subjects were in a sitting position. Across a respiratory rate

range of 6 to 60 breaths per minute (bpm) spanning from slow breathing to respiratory distress scenarios; the optical flow method demonstrated robust performance. When compared to a reference wearable sensor, the system achieved a bias of -0.03 ± 1.38 bpm at 2-meter camera distance and -0.02 ± 1.92 bpm at 0.5-meter distance. These near-zero bias values indicate negligible systematic error, with the system slightly underestimating respiratory rate by an average of only 0.02-0.03 bpm. The standard deviations (1.38 bpm at 2m and 1.92 bpm at 0.5m) reflect measurement variability, showing slightly better consistency at greater distance. However, validation was limited to healthy volunteers in controlled postures and did not evaluate performance during actual respiratory distress conditions characterized by irregular breathing patterns, chest retractions, or paradoxical breathing.

Huang et al. (Huang *et al.*, 2024) introduced a contactless respiratory imaging system that represents a significant advancement in infant respiratory monitoring by extending beyond conventional respiratory rate measurement to analyze thoracoabdominal breathing patterns. Their system employs optical flow methods to generate motion intensity maps that quantify the contribution of different body regions during respiration and phase representations that capture the temporal coordination between chest and abdominal movements. Phase analysis determines whether the chest and abdomen move synchronously (in-phase, expanding and contracting together during normal breathing) or asynchronously (out-of-phase, moving in opposite directions—a pathological pattern indicating respiratory distress). These intensity and phase maps are analyzed using advanced deep learning techniques to classify thoracoabdominal asynchronous respiration. To address the challenge of limited training data, they proposed a novel multiple-expert contrastive learning framework with two main components: (1) contrastive pairing-based data augmentation, which enriches the dataset by symmetrically reversing intensity images and pairing them randomly with phase images from different breathing pattern classes (synchronous versus asynchronous breathing), and (2) multi-expert contrastive optimization, which employs multiple expert networks with prototype-based contrastive learning to ensure consistent learned representations across diverse infant physiology. Clinical validation involving 44 infants across two neonatal intensive care units demonstrated an accuracy of 76.56%,

sensitivity of 70.72%, and specificity of 80.21% for thoracoabdominal asynchronous detection. These results establish the feasibility of contactless detection of pathological parameters critical for early identification of neonatal respiratory compromise.

1.3 RGB-D Camera-Based Contactless Monitoring Methods

Recent advances in consumer-grade RGB-D cameras have motivated the use of depth information for respiratory monitoring. By capturing three-dimensional structural data in addition to color information, depth cameras enable more accurate and robust contactless estimation of respiratory parameters compared to conventional RGB imaging.

Xia et al. (Xia & Siochi, 2012) proposed an early contactless respiratory monitoring system using the Microsoft Kinect sensor, which simultaneous acquires depth and color information. Their approach employed a translation surface on the torso to reduce noise from irregular surface geometries caused by clothing and fabric wrinkles. Subsequently, they amplified the horizontal motion using signal processing techniques to capture enhanced respiratory signal. The system tracked average depth values over time (30 Hz) within a defined thoracic ROI as a measure of chest wall displacement. correlation coefficients between Kinect derived respiratory signals and with standard measurements devices (strain gauge belt) ranged from 0.958 to 0.978, demonstrating strong agreement and applicability.

Benetazzo et al. (Benetazzo, Freddi, Monteriù & Longhi, 2014) employed RGB-D structured light cameras for respiratory monitoring. Unlike Kinect, which generates independent depth images, structured light cameras compute depth values for each RGB pixel, producing spatially aligned RGB-D video sequences. Their system utilized the OpenNI[1] library to automatically detect the ROI, defined by bilateral shoulder landmarks and the torso. Mean distance values within the defined region were computed frame by frame. Respiratory rate was extracted from the resulting distance time graph. The system demonstrated several advantages through automated ROI detection with continuous re-detection capabilities to accommodate patient

[1] https://github.com/OpenNI/OpenNI

movements. Extensive validation under varying conditions, including different clothing types, patient orientations, illumination levels, and movement patterns, yielded a minimum correlation coefficient of 0.92 compared to reference spirometer measurements.

Harte et al. (Harte *et al.*, 2016) developed a chest wall motion system using four Kinect sensors placed around the patient. Unlike previous single-camera systems that only viewed the chest from the front, this multi-sensor setup captured the entire torso in 3D. The system acquired 3D point clouds from all four cameras, which were aligned into a common coordinate system using calibration data. The resulting 3D surface models were then used to compute breathing-related volume changes over time. A two-phase evaluation assessed the system accuracy: static testing using a cardiopulmonary resuscitation mannequin demonstrated a root mean square error of 0.1 liters (0.441% under-prediction) compared to Nikon laser scanner reference measurements. Dynamic testing with simultaneous spirometry in nine cystic fibrosis patients and thirteen healthy volunteers performing relaxed vital capacity maneuvers yielded Pearson's correlation coefficients exceeding 0.8656 for patients and 0.9226 for healthy volunteers (all $p < 0.001$). Statistical analysis revealed significantly lower correlations in the cystic fibrosis group ($p = 0.0082$), attributed to characteristic airflow obstruction.

Simplifying the multi-camera configuration, Transue et al. (Transue, Nguyen, Vu & Choi, 2016) proposed an automated tidal volume monitoring system using a single Microsoft Kinect depth camera positioned frontally to capture the chest surface. The approach introduced a real-time iso-surface reconstruction algorithm to generate three-dimensional deformation states of the chest during breathing. Unlike direct volumetric measurement approaches, the system employed a per patient correlation metric acquired through bayesian network to estimate breathing volume from reconstructed chest surface. The single camera configuration offered practical advantages in system cost, setup simplicity, and reduced computational complexity. However, the frontal viewpoint captured primarily chest wall motion, potentially missing lateral and posterior deformation patterns, and the requirement for per-patient Bayesian network training prevent immediate measurements without prior calibration.

Rehouma et al. (Rehouma *et al.*, 2018) developed a comprehensive 3D imaging system for estimation of tidal volume and respiratory rate during spontaneous breathing in clinical settings. The system utilized two RGB-D cameras positioned at an angle of 45° with one at the bottom right corner of the bed and the other at the top left corner of the bed. Their system captured point-cloud information which was subsequently aligned using precomputed transformation matrices. Three-dimensional thoracic surface meshes were generated through Poisson surface reconstruction (Kazhdan, Bolitho & Hoppe, 2006), followed by frame-by-frame volumetric quantification using Octree-based spatial decomposition. Tidal volume was derived by computing the difference between maximum and minimum volumes across the respiratory cycle. Rehouma et al. (Rehouma, Noumeir, Masson, Essouri & Jouvet, 2019b) proposed a contactless system to detect and quantify thoracoabdominal asynchrony, a critical indicator of respiratory distress. Their system used a single RGB-D camera to capture torso information. The method employed advance computer vision techniques to compute 3D scene flow (Jaimez, Kerl, Gonzalez-Jimenez & Cremers, 2017), segmenting the thoracic and abdominal regions based on the direction of motion vectors. Displacement measurements for the respective regions were obtained using the Euclidean distance method, with time-distance graphs enabling quantification of asynchronous breathing patterns.

Nazir et al. (Nazir *et al.*, 2021) proposed a novel contactless solution for real-time respiratory monitoring and function assessment of intensive care units patients using RGB-D sensors. The system analyzed chest wall morphological changes to estimate multiple respiratory parameters simultaneously. A key innovation was the integration of a deep neural network model, trained on the COCO dataset, for automatic torso detection, eliminating the need for manual region of interest selection. The system computed respiratory parameters including tidal volume and respiratory rate by analyzing temporal changes in chest wall surface topology. Comprehensive evaluation was conducted on both a controlled mannequin setup and 16 mechanically ventilated patients in the ICU of Brest University Hospital, comprising 216 recordings in total. The clinical validation demonstrated the system's capability to function in realistic ICU environments.

The deep learning based torso detection achieved 88.8% accuracy, enabling robust automated operation.

Addison et al. (Addison *et al.*, 2022) developed a contactless depth camera system for continuous monitoring of respiratory parameters in hospitals. The system tracked distance changes between the camera and the patient's torso surface to measure respiratory rate and volume. They tested the system against ventilator measurements using one volunteer with controlled breathing patterns. Results showed minimal error for respiratory rate (bias: -0.02 bpm, RMSE: 0.51 bpm, $R = 0.98$, $p < 0.001$) and tidal volume (bias: -0.21 L, RMSE: 0.23 L, $R = 0.92$, $p < 0.001$). In a follow-up study, Addison et al. (Addison *et al.*, 2023) tested an Intel RealSense™ D415 depth camera for respiratory rate monitoring in more challenging conditions. Seven healthy subjects performed controlled breathing at rates from 4 to 40 breaths per minute. The study collected 553 recordings across different postures, positions, lighting levels (including complete darkness), and bed coverings. The system achieved high accuracy with error below 1.0 breath per minute when compared to capnography. It maintained accuracy even with patient movement, changing light conditions, and thick bed coverings. However, the study only tested healthy volunteers with controlled breathing. Performance on sick patients or abnormal breathing patterns was not evaluated.

Chavernac et al. (Chavernac *et al.*, 2025) developed a real-time, contactless respiratory monitoring system tailored for pediatric intensive care units, utilizing the Azure Kinect depth camera. The system automatically tracked variations in thoracic volume to derive a wide range of ventilatory parameters, including tidal volume, respiratory rate, minute ventilation, inspiratory and expiratory times, the I:E ratio, and peak flow rates. A notable innovation was the use of infrared-based region of interest detection via YOLO-OBB, which enabled robust operation even in complete darkness and demonstrated superior performance relative to RGB-based detection (mAP@50-95: 0.93 vs. 0.77). Additionally, pixel-wise 3D volume computation yielded a mean absolute error below 5% for tidal volume measurements. Validation involved one healthy adult, showing a Pearson correlation of 0.995 with spirometry, and one critically ill non-intubated pediatric patient, with errors of 1.5% for minute expiratory volume and 2% for tidal volume.

Table 1.2 Summary of RGB and RGB-D camera-based respiratory monitoring studies

Reference	Method	Camera Type	Parameters
Bai *et al.* (2010)	Temporal Differencing	2× RGB	RR, Apnea
Xia & Siochi (2012)	Camera to Object Distance	1× Kinect	RR
Aarts *et al.* (2013)	Temporal Variation	1× RGB	HR
Bartula *et al.* (2013)	Vertical Profile Projection	1× RGB	RR
Benetazzo *et al.* (2014)	Camera to Object Distance	1× Structured Light	RR
Tarassenko *et al.* (2014)	Auto-Regressive Model	1× RGB	RR, HR
Al-Naji & Chahl (2016)	Eulerian Video Magnification	1× RGB	RR
Transue *et al.* (2016)	Distance Measurement	1× Kinect	Vt, RR
Harte *et al.* (2016)	Surface Reconstruction	4× Kinects	Vt, RR
Rehouma *et al.* (2018)	Surface Reconstruction, Octree	2× Kinects	Vt, RR
Massaroni *et al.* (2019)	Color Intensity Variations	1× RGB	RR
Rehouma *et al.* (2019b)	3D Scene Flow	1× Kinect	TAA
Mateu-Mateus *et al.* (2020)	Dense Optical Flow	1× RGB	RR
Brieva *et al.* (2020)	Motion Magnification & CNN	1× RGB	RR
Romano *et al.* (2021)	Optical Flow, Artifact Removal	1× RGB	RR, Apnea
Nazir *et al.* (2021)	Deep Neural Network	1× Kinect	Vt, RR
Addison *et al.* (2022)	3D Volume Computation	1× Kinect	RR, Vt
Addison, Antunes, Montgomery, Smit & Borg (2023)	3D Volume Computation	1× Kinect	RR
Huang *et al.* (2024)	Optical Flow, Deep Learning	1× RGB	TAA
Chavernac *et al.* (2025)	3D Volume Computation	1× Kinect	Vt, RR, MV

RR: Respiratory Rate | Vt: Tidal Volume | TAA: Thoracoabdominal Asynchrony | MV: Minute Ventilation

Research GAP

As summarized in Table 1.2, previous studies utilizing RGB and RGB-D cameras have successfully measured fundamental respiratory parameters including respiratory rate, tidal volume, thoracoabdominal asynchrony, and heart rate in a contactless manner. These vision-based approaches offer significant advantages over traditional contact-based monitoring systems in pediatric intensive care settings. Notably, a single camera-based system can simultaneously capture multiple physiological parameters that would otherwise require several separate contact-based sensors, thereby reducing system complexity and patient burden. These methods typically employ advanced computer vision techniques, particularly optical flow, combined with signal processing approaches to extract respiratory signals. The contactless nature of these systems enables continuous monitoring even during patient movement while eliminating risks associated with adhesive sensors, such as skin breakdown and device-related complications.

Despite these advances and the extensive automation achieved in measuring individual respiratory parameters, a critical gap remains: the automated detection of acute respiratory distress through analysis of chest wall deformation patterns is largely understudied. While recent research (Rehouma *et al.*, 2019b; Huang *et al.*, 2024) has explored thoracoabdominal asynchrony assessment, these investigations have been limited to controlled settings with healthy subjects or mannequin simulations, lacking validation in clinical populations with actual respiratory pathology. This gap is particularly significant given the potential of chest wall deformation analysis to provide early indicators of respiratory deterioration in critically ill patients.

1.4 Video Analysis Algorithms

Video understanding remains a fundamental challenge in computer vision, requiring models to capture both spatial features within frames and temporal dynamics across frame sequences. Prior to deep learning, conventional handcrafted methods dominated the field. These methods employed optical flow algorithms (Lucas-Kanade, Horn-Schunck), dense trajectories, Histogram of Oriented Gradients (HOG) and Scale-Invariant Feature Transform (SIFT) descriptors and

Space-Time Interest Point (STIP) detectors encoded via bag-of-words representations for action recognition tasks (Mahbub, Imtiaz & Ahad, 2011; Wang & Schmid, 2013; Burghouts & Schutte, 2013; Wang, Kläser, Schmid & Liu, 2013; Zhang, Tsoi & Lo, 2014; Yang & Kurita, 2014). While these approaches achieved notable success on constrained benchmarks, they suffered from limited representational capacity and required extensive domain expertise for feature engineering.

The advent of deep learning revolutionized the computer vision field through automatic feature learning from large-scale labeled datasets via backpropagation. 2D CNNs have demonstrated outstanding performance in image recognition by effectively learning spatial representations from images. Nevertheless, their frame-by-frame processing approach fundamentally limits their ability to capture temporal dependencies and motion information across video sequences.

Ji et al. (Ji, Xu, Yang & Yu, 2012) introduced the 3D CNN architecture that expand 2D kernel to the temporal dimension by applying 3D kernels of size $T \times k \times k$ (temporal depth $\times$ spatial height $\times$ width) to video volumes. The network processes input clips through hierarchical 3D convolutional layers where each kernel slides across both spatial and temporal dimensions, computing dot products between filter weights and local spatiotemporal patches to extract features such as moving edges, corner trajectories, and texture dynamics. Following each convolutional layer, 3D max pooling operations with $1 \times 2 \times 2$ kernels provide local spatial invariance while reducing computational costs and expanding receptive fields for deeper layers.

Building upon this foundation, the research community has developed increasingly sophisticated architectures that address the computational challenges and modeling requirements of video understanding. This review examines the video analysis algorithms, focusing on six representative architectures specifically used for this study: 2D CNN+LSTM (Veeriah, Zhuang & Qi, 2015), R(2+1)D (Tran *et al.*, 2018), SlowFast (Feichtenhofer, Fan, Malik & He, 2019), S3D (Xie, Sun, Huang, Tu & Murphy, 2018), X3D (Feichtenhofer, 2020), and the advanced Swin-3D (Liu *et al.*, 2022) hybrid models.

Tran et al. (Tran *et al.*, 2018) proposed the R(2+1)D action recognition network, which factorizes standard 3D convolutions layer into sequential 2D spatial convolutions ($1 \times d \times d$) followed by 1D temporal convolutions ($t \times 1 \times 1$) layer. This approach effectively decomposes a $3 \times 3 \times 3$ kernel into a $1 \times 3 \times 3$ spatial operation that processes each frame independently to extract spatial features (edges, textures, object parts), followed by ReLU activation. A $3 \times 1 \times 1$ temporal operation then aggregates information across 3 consecutive time steps, followed by another ReLU activation. This factorization doubles the number of nonlinearities compared to standard 3D convolutions while maintaining similar parameter counts (by adjusting the intermediate channel dimension $M \approx 2.25C$ to match parameters). It enables increased model expressiveness, easier optimization through simpler gradient flow, and implicit regularization through structured spatiotemporal interaction. This architecture achieves 76% top-1 accuracy on the Kinetics-400 dataset.

Xie et al. (Xie *et al.*, 2018) introduced S3D (Separable 3D CNNs), which decomposes 3D convolutions into spatial and temporal components similar to R(2+1)D (Tran *et al.*, 2018). However, it advocates for a "top-heavy" design techniques where temporal modeling capacity increases in deeper layers. Unlike uniform factorization approaches, S3D employs learned spatiotemporal separability. Early layers may use full $3 \times 3 \times 3$ convolutions for low-level features while deeper layers use factorized operations with progressively increasing temporal receptive fields. The architecture incorporates temporal self-gating mechanisms that apply sigmoid activations to recalibrate features in the temporal dimension, analogous to squeeze-and-excitation but operating on time. This top-heavy design reduces computation in early layers, enabling higher spatial resolution inputs or longer video clips. This architecture achieves 68% top-1 accuracy on the Kinetics-400 dataset.

Tran et al. (Tran, Wang, Torresani & Feiszli, 2019) proposed channel-separated networks (CSN) that factorize 3D convolutions into pointwise $1 \times 1 \times 1$ convolutions for channel mixing followed by depthwise $3 \times 3 \times 3$ convolutions for per-channel spatiotemporal processing within ResNet-style bottleneck blocks. Two variants exist: ip-CSN (interaction-preserved) adds a post-depthwise $1 \times 1 \times 1$ convolution to maintain channel interactions, while ir-CSN (interaction-reduced) limits

interactions through bottleneck projections only. This factorization achieves 2–3× reduction in FLOPs and parameters compared to standard 3D ResNets while providing regularization benefits. This is evidenced by lower training accuracy but higher test accuracy. Additionally, this network employs temporal striding to progressively reduce temporal resolution throughout the network, enabling the model to capture long-range temporal dependencies in extended video sequences. This architecture achieves 77% top-1 accuracy on the Kinetics-400 dataset.

Feichtenhofer et al. (Feichtenhofer *et al.*, 2019) introduced SlowFast networks, a dual-pathway architecture inspired by biological visual systems. The Slow pathway operates at low temporal resolution (2–4 FPS) with high channel capacity (e.g., 256 channels) to capture spatial semantics using non-degenerate 3D convolutions mainly in deeper stages (res4, res5). The Fast pathway processes frames at high temporal rates (16–32 FPS, $\alpha = 8\times$ faster via temporal stride ratio) with reduced channel capacity ($\beta = 1/8$). Temporal information is retained throughout the network (no temporal downsampling until global pooling). Pathways exchange information through lateral $5 \times 1 \times 1$ time-strided convolutions. This architecture achieves 79% top-1 accuracy on the Kinetics-400 dataset.

Feichtenhofer et al. (Feichtenhofer, 2020) presented the X3D video action classification model, which employs progressive network expansion from a minimal 2D baseline. It systematically scales six dimensions: temporal duration, frame rate, spatial resolution, network width, network depth, and bottleneck width. The expansion process evaluates each dimension independently using a multiplicative factor γ to optimize accuracy-efficiency trade-offs. X3D incorporates inverted residual blocks with spatiotemporal depthwise separable convolutions (separating $1 \times 1 \times 1$ expansion, $d \times k \times k$ depthwise, and $1 \times 1 \times 1$ projection), squeeze-and-excitation channel attention, and swish activations. The architecture uses 3D convolutions with varying kernel sizes optimized per layer and employs both spatial and temporal pooling throughout the network. This progressive expansion strategy generates a family of models from X3D-XS (70.4% top-1 accuracy) to X3D-XL (81.9% top-1 accuracy) on Kinetics-400.

Liu et al. (Liu *et al.*, 2022) extended the image Swin Transformer to video understanding, by replacing 2D windows with hierarchical 3D spatiotemporal self-attention. Initially, they applied a 3D convolution layer with kernel size of $2 \times 7 \times 7$ and stride of $2 \times 4 \times 4$ to map the input videos to low-level features. Subsequently, the architecture uses 3D shifted window partitioning, dividing clips into local regions for self-attention with linear complexity $O(T \times H \times W)$. Shifted partitioning alternates across blocks for cross-window connections. Four stages progressively downsample spatial resolution via $1 \times 2 \times 2$ patch merging, with 3D relative position biases encoding spatiotemporal relations. The top variant of this architecture achieves 82% top-1 accuracy on the Kinetics-400 dataset.

While the aforementioned architectures represent the key models examined in this study, many more architectures have been presented in the literature, each contributing unique innovations to video understanding and action recognition tasks.

CHAPTER 2

AUTOMATED DETECTION OF ACUTE RESPIRATORY DISTRESS USING TEMPORAL VISUAL INFORMATION

Wajahat Nawaz[1] , Philippe Jouvet[2] , Rita Noumeir[1]

[1] Department of Electrical Engineering, École de Technologie Supérieure, 1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] Research Center at CHU Sainte-Justine Hospital, University of Montreal, 3175 Chem. de la Côte-Sainte-Catherine, Montréal, Québec, Canada H3T 1C5

Article published in «*IEEE Access* » in October 2024

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 based 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.

2.1 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 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) (Edwards *et al.*, 2013a). In severe cases, ARD has a high mortality rate, with about 40% of patient deaths resulting from the condition (Ramji, Hafiz, Altaq, Hussain & Chaudry, 2023).

ARD is a life-threatening condition, as prolonged and excessive use of accessory muscles may progress to respiratory failure and subsequent cardiopulmonary arrest (Jaeger *et al.*, 2019). 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 (Mirabile *et al.*, 2023). 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 (National Heart, Lung, and Blood Institute (NHLBI), 2025). Among these indicators,

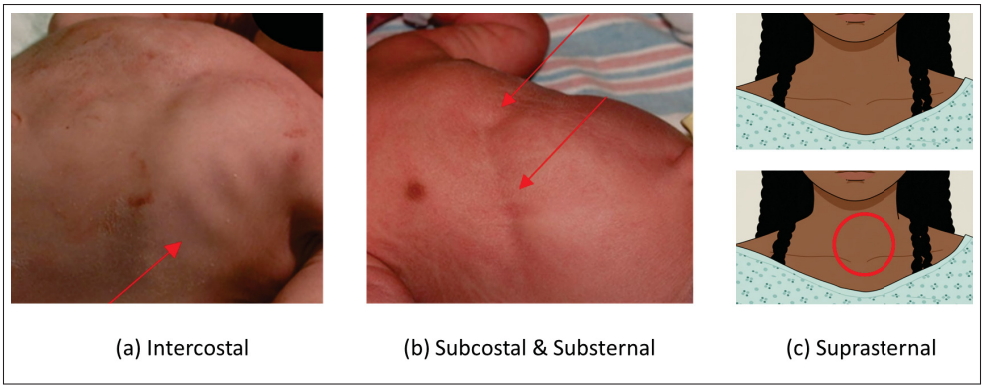

Figure 2.1 Examples of chest retraction signs and their potential locations. Image credit: (a, b) Photos by Janelle Aby, MD, Stanford Medicine Newborn Nursery (Stanford Medicine Newborn Nursery, 2025)

the identification of chest retraction is a crucial, frequently utilized, non-invasive method for evaluating the severity of ARD, especially in the PICU (McCollum & Ginsburg, 2017). Figure 2.1 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 (Bai *et al.*, 2010; Xia & Siochi, 2012; Benetazzo *et al.*, 2014; Lee, Pathirana, Evans & Steinfort, 2015; Rehouma, Noumeir, Essouri & Jouvet, 2019a; Cheng *et al.*, 2023), sleep apnea event detection (Feng, Qin, Wu, Pan & Liu, 2020; Shen, Qin, Wei & Liu, 2021; Bahrami & Forouzanfar, 2022; Chen, Ma, Gao & Fan, 2023), tidal volume estimation (Transue *et al.*, 2016; Rehouma *et al.*, 2018; Yuthong, Duangsoithong, Booranawong & Chetpattananondh, 2019), chest deformation assessment (Rehouma *et al.*, 2019b; Di Tocco *et al.*, 2020; Ottaviani *et al.*, 2022).

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 2.2. 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).

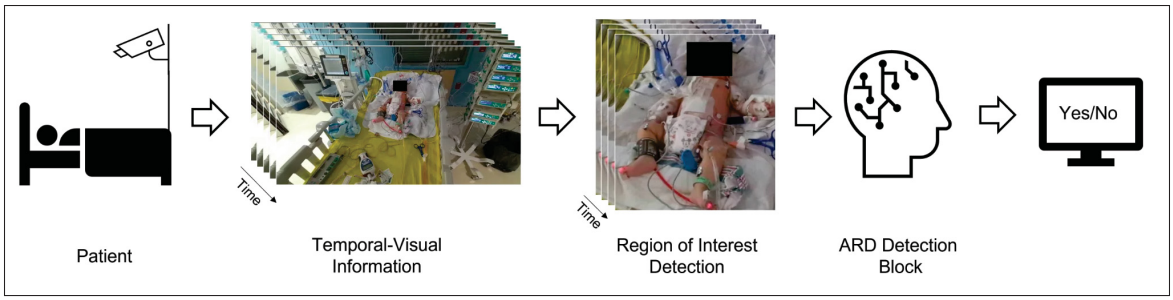

Figure 2.2 High level overview of our proposed acute respiratory distress detection system. 1). High-definition camera captures the temporal visual information and 2). ROI is extracted within the video then 3). fed into the ARD detection block

The rest of the paper is organized as follows. Section 2.2 presents the related work and identifies the research gap. Section 2.3 details the data acquisition and labeling process. Section 2.4 outlines the proposed methodology, including data processing steps such as temporal and spatial region of interest extraction, and the ARD detection system architecture. Section 2.5 covers the implementation details, experimental results, cross-model evaluation, and qualitative analysis. Section 2.6, discusses the results and limitations of the work. Finally, Section 2.7 concludes the paper and outlines potential future work.

2.2 Literature Review

Traditionally, the respiratory rate was estimated by counting the respiration cycles (inspiration and expiration) for one minute. However, with the advancement of technology, Bai et al. (Bai *et al.*, 2010) 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 et al. (Xia & Siochi, 2012) 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. (Benetazzo *et al.*, 2014) advanced the work of Xia et al. (Xia & Siochi, 2012) 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. (Lee *et al.*, 2015) employed a Microwave Doppler Radar (MDR) sensor to estimate the distance-to-time graph for respiratory rate estimation. Mateu et al. (Mateu-Mateus *et al.*, 2020) 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. (Cheng *et al.*, 2023) 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. (Harte *et al.*, 2016) and Transue et al. (Transue *et al.*, 2016) 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. (Rehouma *et al.*, 2018) 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 (Kazhdan *et al.*, 2006) 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 (Rehouma *et al.*, 2019a).

Rehouma et al. (Rehouma *et al.*, 2019b) 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 (Jaimez *et al.*, 2017) 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. (Di Tocco *et al.*, 2020) 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. V. Ottaviani et al. (Ottaviani *et al.*, 2022) 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 (Bai *et al.*, 2010; Xia & Siochi, 2012; Benetazzo *et al.*, 2014; Lee *et al.*, 2015; Rehouma *et al.*, 2019a; Cheng *et al.*, 2023) and tidal volume estimation (Transue *et al.*, 2016; Rehouma *et al.*, 2018; Yuthong *et al.*, 2019), while very few discuss thoraco-abdominal asynchrony (Rehouma *et al.*, 2019b; Di Tocco *et al.*, 2020; Ottaviani *et al.*, 2022). 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.

2.3 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 2.3. 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.

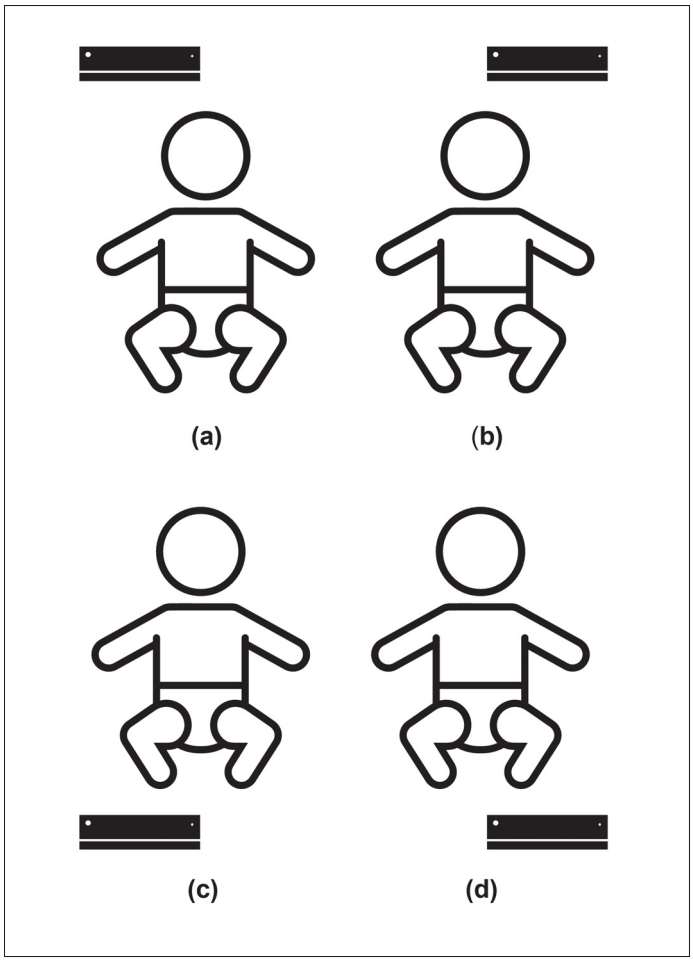

Figure 2.3 Available camera position in the pediatric intensive care units: top a). left and b). right, and bottom c). left and d). right

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.

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 (Silverman & Andersen, 1956), 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) retractions. Table 2.1 shows the detailed statistics of acute respiratory distress patients and the associated retraction signs.

2.4 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

Table 2.1 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

Algorithm 2.1 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 ≥ 0.5:
        Return 1 (patient has *ARD*)
    Else:
        Return 0 (patient has no *ARD*)
Until the input video is no longer available

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 2.1.

2.4.1 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.

2.4.1.1 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

Table 2.2 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

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.

2.4.1.2 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 (Zhu *et al.*, 2022), 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 2.2 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.

2.4.2 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 (Klaser, Marszałek & Schmid, 2008) and optical flow (Mahbub *et al.*, 2011) techniques. However, recent advancements in deep convolutional neural networks (CNNs) have revolutionized vision-related tasks, including image classification (Iandola *et al.*, 2016), object localization and segmentation (He, Gkioxari, Dollár & Girshick, 2017), and action recognition (Ji *et al.*, 2012). 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 2.4 shows the framework of a *2D-CNN + RNN* based ARD detection system. This approach combines the strengths of *2D-CNNs* and *RNNs* (Yue-Hei Ng *et al.*, 2015) 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) (Hochreiter & Schmidhuber, 1997) or GRU (gated recurrent unit)(Dey & Salem, 2017). Initially, *RNNs* were designed to deal with time series data such as text classification (Liu & Guo, 2019), sound classification (Bubashait & Hewahi, 2021). Later on, researchers employed it for action recognition task (Yue-Hei Ng *et al.*, 2015; Ullah, Ahmad, Muhammad, Sajjad & Baik, 2017; Savran Kızıltepe, Gan & Escobar, 2021) 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 them computationally efficient during inference 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* (Tran, Bourdev, Fergus *et al.*, 2015) 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 (Tran *et al.*, 2015, 2019; Xie *et al.*,

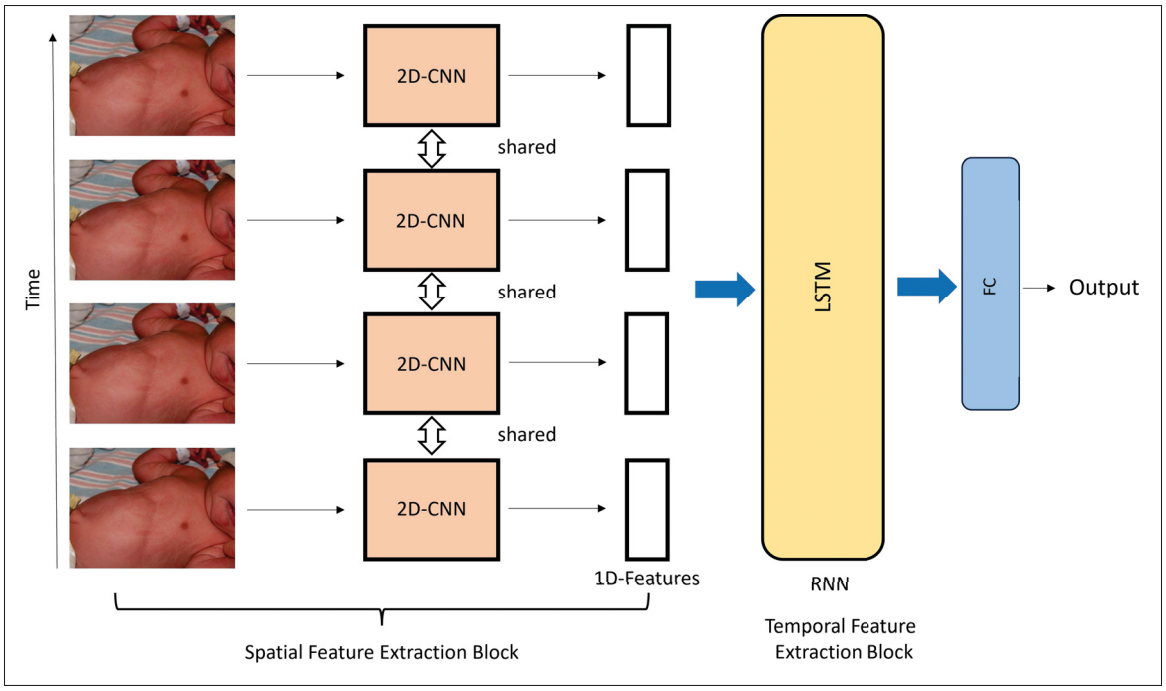

Figure 2.4 A 2D convolution neural network based respiratory distress (ARD) detection system. 1) Spatial Features Extraction Block: This block focuses on extracting spatial features from each video frame independently. 2) Temporal Feature Extraction Block: It extracts the temporal features that capture the relationships between frames. A recurrent neural network, specifically LSTM (Long Short-Term Memory), is employed to extract the temporal features

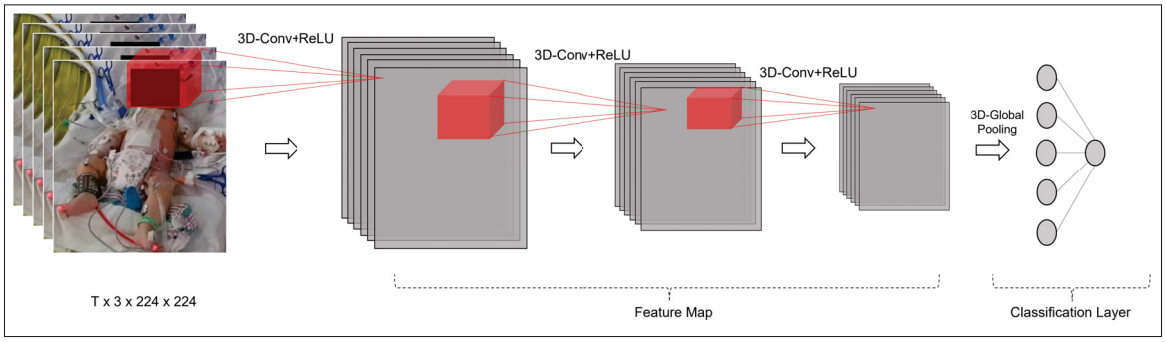

Figure 2.5 3D Convolution neural network based Acute Respiratory Distress detection network

2018; Feichtenhofer *et al.*, 2019). Figure 2.5 shows 3D convolutions neural networks based ARD detection system.

2.5 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.

2.5.1 Implementation Details

For model training and testing, we divide 30-second-long patient videos into smaller video clips, as discussed in Section 2.4.1.2. 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.

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 (Szymański & Kajdanowicz, 2017), 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.

2.5.2 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) (Tran *et al.*, 2019), recognized for its superior performance in human action recognition (HAR) tasks on the Kinetics-400 dataset. We fine-tuned the $CSN - R101$ originally trained on the Kinetics-400 HAR dataset on the ARD dataset by modifying the classification head rather than training the model from scratch.

2.5.2.1 Baseline

Table 2.3 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

Table 2.3 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**

(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.

2.5.2.2 Spatial segmentation

Table 2.3 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 2.4 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

2.5.2.3 Frame Sampling Rate

The optimal frame sampling rate (step size) depends on the specific information the model aims to detect (Diba *et al.*, 2017). 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 2.4, 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 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.

2.5.2.4 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 2.5 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 2.4.1.2. The results for a video clip duration of 3.2 seconds are presented in the first row of Table 2.5. 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 2.5. 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.

Table 2.5 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

2.5.3 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 2.6. 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 over-fitting. 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 2.6, 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 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.

2.5.4 Qualitative Analysis

Figure 2.6 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.

2.6 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.

Table 2.6 Performance comparison of 2D-CNN+LSTM and 3D-CNN architectures

Model type		**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**	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**	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

Footnote: The learning rates are set to 1×10^{-3} for 2D-CNNs+LSTM and 5×10^{-5} for 3D-CNN.

We use stochastic gradient descent (SGD) optimizer with a momentum of 0.90.

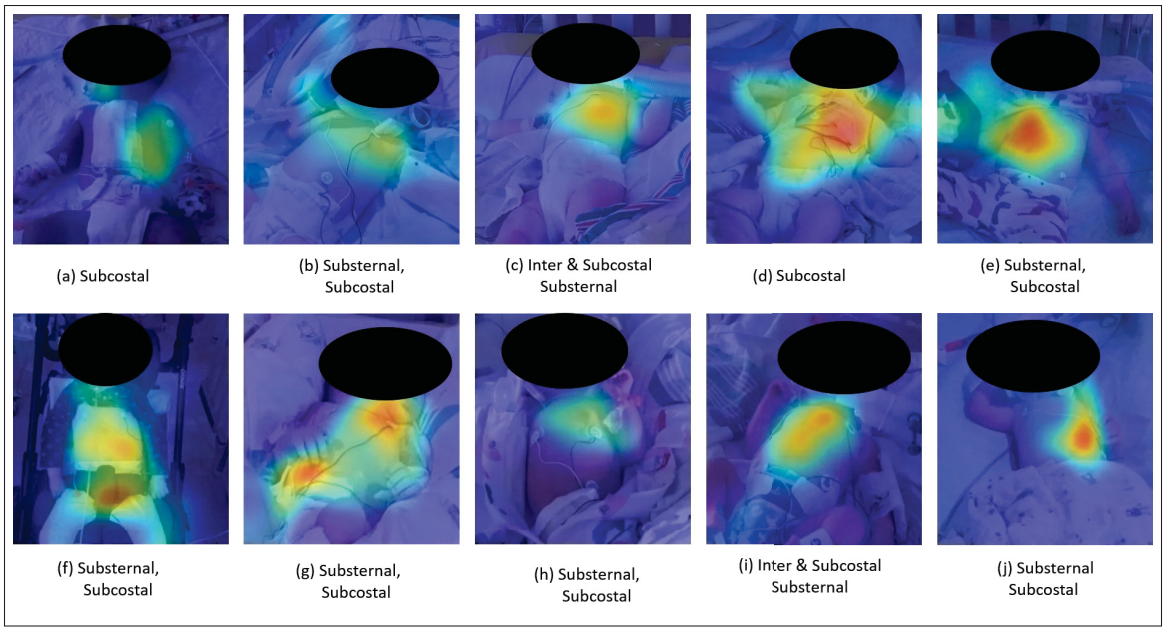

Figure 2.6 Qualitative results of an acute respiratory distress detection model using class activation maps (CAM) reveal that in the model has learned problem-specific features, concentrating exclusively on the torso region of the patients. Nevertheless, in some cases such as (a), (b), (f), and (g), it also learns irrelevant features, focusing on non-interested regions

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.

2.7 CONCLUSIONS & 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.

CHAPTER 3

ACUTE RESPIRATORY DISTRESS IDENTIFICATION VIA MULTI-MODALITY USING DEEP LEARNING

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

[1] Department of Electrical Engineering, École de Technologie Supérieure, 1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] Research Center at CHU Sainte-Justine Hospital, University of Montreal, 3175 Chem. de la Côte-Sainte-Catherine, Montréal, Québec, Canada H3T 1C5

Article published in « *Applied Sciences* » in January 2025

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.

3.1 Introduction

Acute respiratory distress (ARD) is a leading cause of infant admissions to the pediatric intensive care unit (PICU) (Edwards, Kotecha & Kotecha, 2013b). This life-threatening condition is characterized by insufficient oxygen saturation levels in the bloodstream, often resulting from underlying lung diseases (Diamond *et al.*, 2025). 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 (Edwards *et al.*, 2013b).

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 (Taussig & Landau, 2008). 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 (Chen, Zhu, Zhang, Wu & Wang, 2019; Rehouma *et al.*, 2019a; Fiedler, Rapczyński & Al-Hamadi, 2020; Cheng *et al.*, 2023; Fiedler, Werner, Rapczyński & Al-Hamadi, 2023), heart rate estimation (Gupta, Bhowmick & Pal, 2017; Pilz, Zaunseder, Krajewski & Blazek, 2018; Sabokrou, Pourreza,

Li, Fathy & Zhao, 2021; Gao, Wu, Geng & Lv, 2022; Su *et al.*, 2024), tidal volume estimation (Yuthong *et al.*, 2019; Hurtado, Chavez, Mansilla, Lopez & Abusleme, 2020; Addison *et al.*, 2022; Rehouma *et al.*, 2018), and thoracic-abdominal asynchrony (TAA) assessment (Rehouma *et al.*, 2019b; Di Tocco *et al.*, 2020; Ottaviani *et al.*, 2022). 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 (Nawaz, Jouvet & Noumeir, 2024), 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 3.2 reviews relevant literature on current techniques for analyzing respiratory parameters and methods for multi-modality feature fusion. Section 3.3 provides an overview of the proposed model, detailing the pre-processing techniques, feature extraction module, and multi-modality feature fusion. Section 3.4 describes the database, implementation details, and presents the experimental results. Finally, Section 3.6 discusses the findings and provides concluding remarks.

3.2 Related Work

3.2.1 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. (Mateu-Mateus *et al.*, 2020) 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. (Rehouma *et al.*, 2018) 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. (Rehouma *et al.*, 2019b) 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. (Ottaviani *et al.*, 2022) developed a contactless method utilizing depth cameras to monitor infants' breathing patterns and thoracoabdominal asynchronous movements. Nawaz et al. (Nawaz *et al.*, 2024) 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.

3.2.2 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 (Merler, Ratha, Feris & Smith, 2019; Buolamwini & Gebru, 2018) and lighting conditions (Adjabi, Ouahabi, Benzaoui & Taleb-Ahmed, 2020), 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. (Khalid & Yu, 2018) proposed a multi-modal three-stream fusion network, drawing inspiration from the success of two-stream fusion networks (Simonyan & Zisserman, 2014; Feichtenhofer, Pinz & Zisserman, 2016). This approach incorporates RGB spatial information, dense optical flow (temporal) data, and pose features to enhance model performance. Similarly, Islam et al. (Islam & Iqbal, 2020) 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. (Das, Sharma, Dai, Bremond & Thonnat, 2020) designed an attention mechanism specifically to fuse spatial-temporal features with pose features, aiming to enhance the understanding of human actions. Joze et al. (Joze, Shaban, Iuzzolino & Koishida, 2020)

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. (Hu, Zheng, Pan, Lai & Zhang, 2018) utilized a bilinear pooling layer to effectively combine multi-modal features, further enhancing overall model efficacy.

Xu et al. (Weiyao, Muqing, Min & Ting, 2021) 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. (Kini, Fleischer, Dave & Shah, 2023) adopted an ensemble modeling strategy to leverage multi-modal information, achieving first place in the ICIAP-W 2023 challenge. Numerous other fusion (Wang, Hu, Lai, Zhang & Zheng, 2019; Popescu, Mocanu & Cramariuc, 2020; Keskes & Noumeir, 2021; Khalid, Ghadi, Gochoo, Jalal & Kim, 2021) schemes have been proposed to effectively combine multi-modal information, highlighting the growing interest and research in this area.

3.3 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 3.1.

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

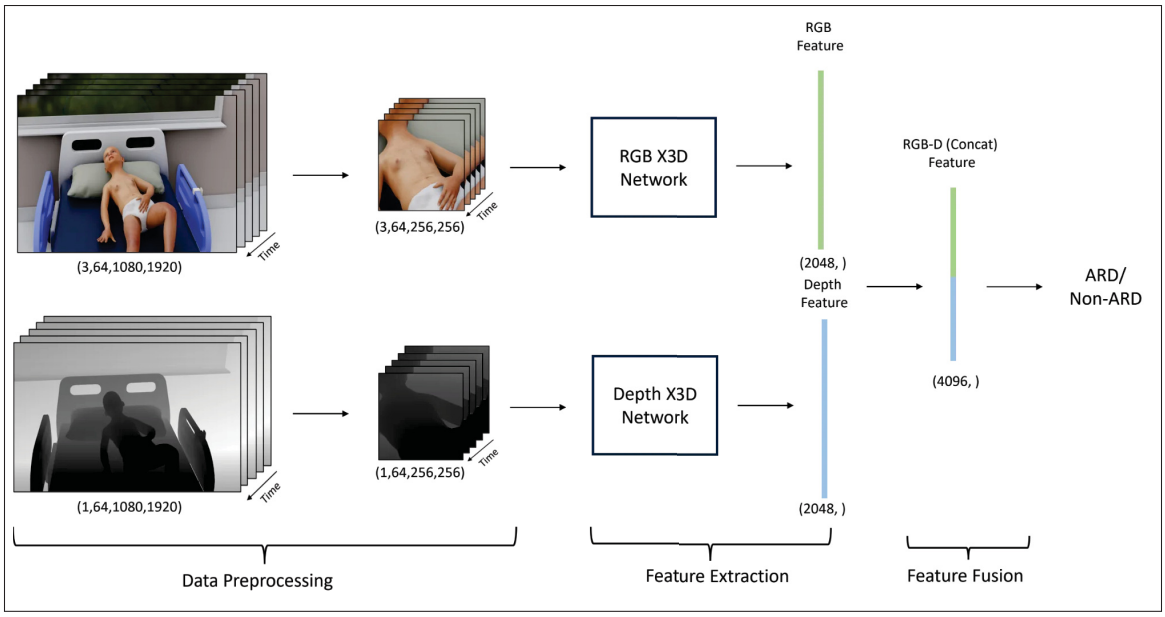

Figure 3.1 Illustration of the proposed network architecture for detecting acute respiratory distress, featuring the integration of RGB and depth temporal visual data through identical 3D convolutional neural networks

discuss the feature fusion techniques used to effectively combine the extracted features from both modalities, enhancing the model's overall performance.

3.3.1 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 (Nawaz *et al.*, 2024) 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.

3.3.2 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 (Zhou, Park & Koltun, 2018), 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 3.2. This step is inspired by previous studies (Tsou, Lee, Hsu & Chang, 2020; Weiyao *et al.*, 2021; Nawaz *et al.*, 2024; Ouzar, Djeldjli, Bousefsaf & Maaoui, 2023), 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 3.1.

Algorithm 3.1 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 foreach
Return normalized depth video D_{norm};

3.3.3 Feature Extraction Module

We use two X3D (Feichtenhofer, 2020) (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.

3.3.4 Feature Fusion Module

Numerous feature fusion techniques have been proposed, such as bilinear pooling (Hu *et al.*, 2018) and self-attention networks (Weiyao *et al.*, 2021). However, these methods often involve

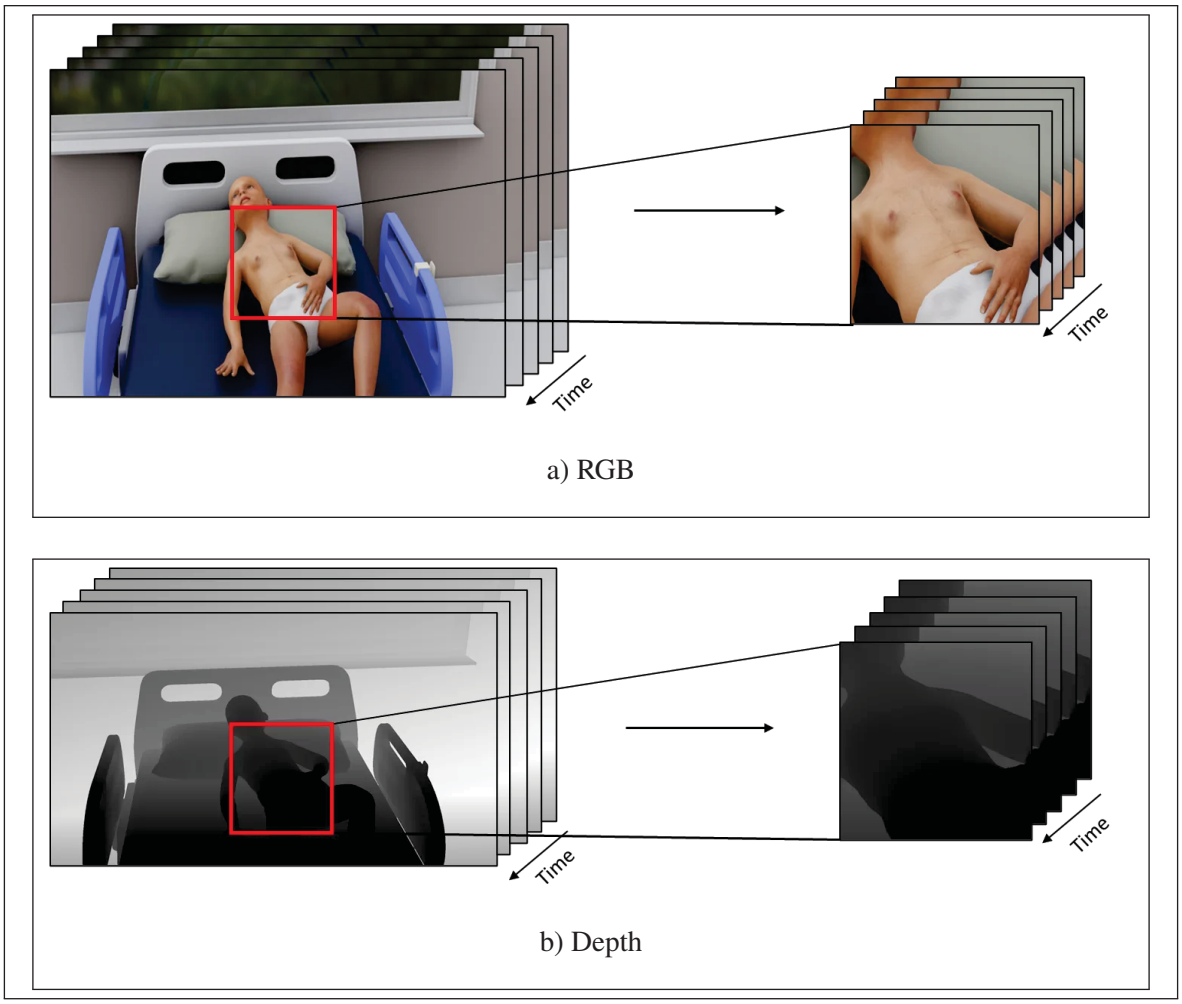

Figure 3.2 RGB-D videos' cropping: (a) RGB and (b) depth

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 (Weiyao *et al.*, 2021).

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 3.1 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.

3.4 Experimental Analysis

3.4.1 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 (Hedstrom, Gove, Mayock & Batra, 2018), 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 3.3. 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.

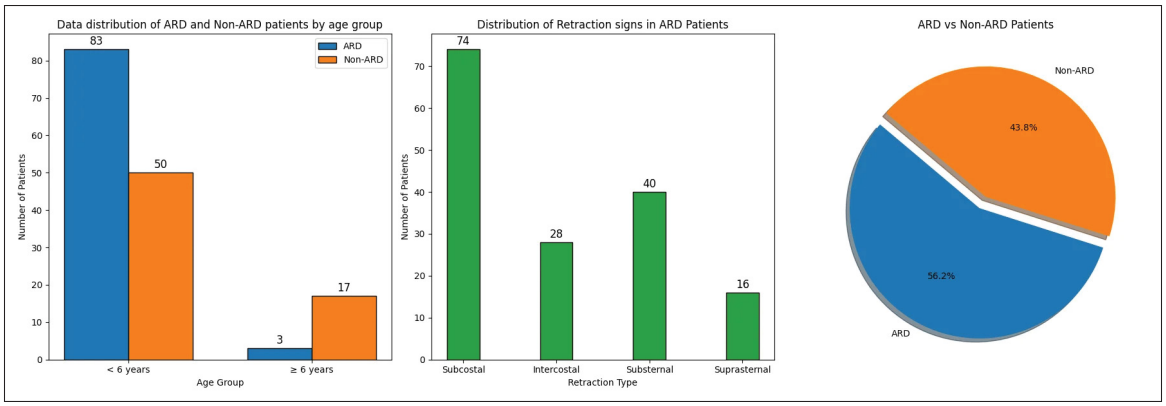

Figure 3.3 Data distribution of ARD and non-ARD patients categorized by age group, retraction type, and overall totals

3.4.2 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 (Szymański & Kajdanowicz, 2017), 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 256×256 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 3.3–3.3.2. 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 224×224, 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.

3.4.3 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.

3.4.4 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.

3.4.4.1 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 3.1 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 3.1 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	Modality	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	0.950	0.720	**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**	1.000	0.750	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	0.700	0.793	0.840	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.

3.4.4.2 RGB-D Acute Respiratory Distress Detection

The experimental results in Table 3.1 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

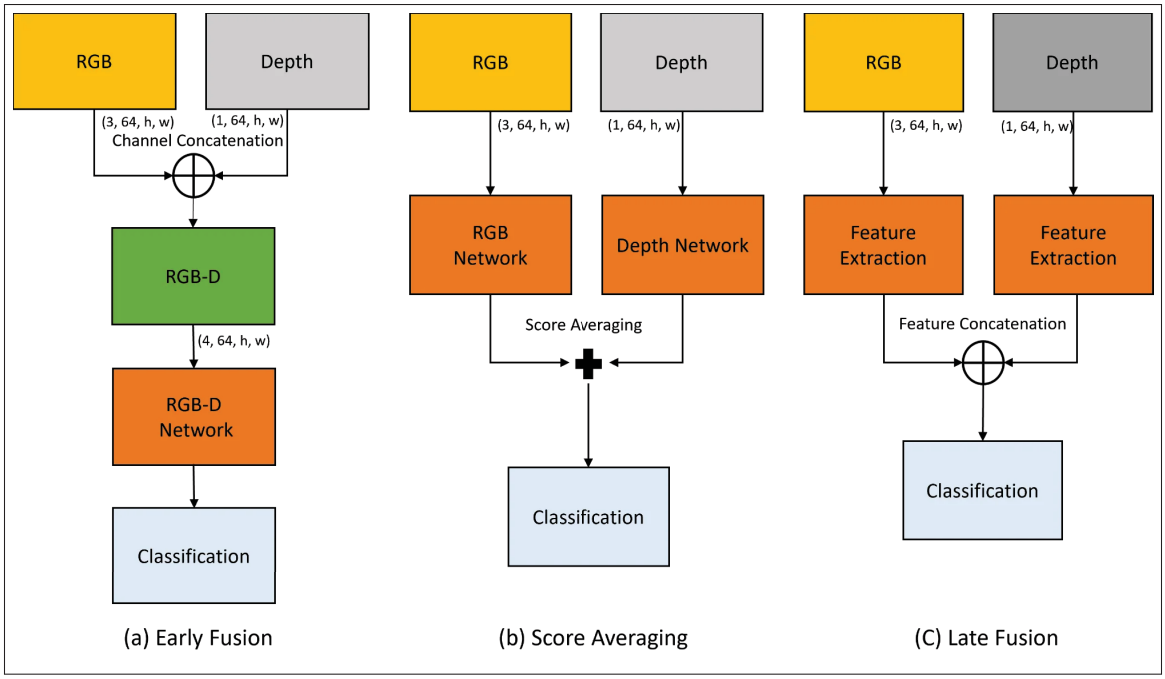

Figure 3.4 Block diagram illustrating the various types of multi-modality fusion schemes: (**a**) early fusion, where input modalities are combined at the input level; (**b**) score averaging, where individual modality predictions are averaged; (**c**) late fusion, where features are combined after independent processing of each modality w/o base-model freezing

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 3.4.

3.4.4.3 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

Table 3.2 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	**Accuracy**			**Precision**			**Recall**			**F_1 Score**		
Method	**Min**	**Avg**	**Max**	**Min**	**Avg**	**Max**	**Min**	**Avg**	**Max**	**Min**	**Avg**	**Max**
Baseline	0.783	0.822	0.870	0.818	0.872	0.950	0.720	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

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 3.2. 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.

3.4.4.4 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 3.2. 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.

3.4.4.5 Performance Analysis Across Age Groups

For this analysis, we group the dataset into two age groups: 1 (<6 years) and 2 (≥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 ≥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 3.3 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 3.3 Model performance across age groups (1: <6 years, 2: ≥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

3.5 Discussion

The experimental results highlight the importance of using multi-modality data for detecting ARD. Figure 3.5 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 3.1). 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.

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.

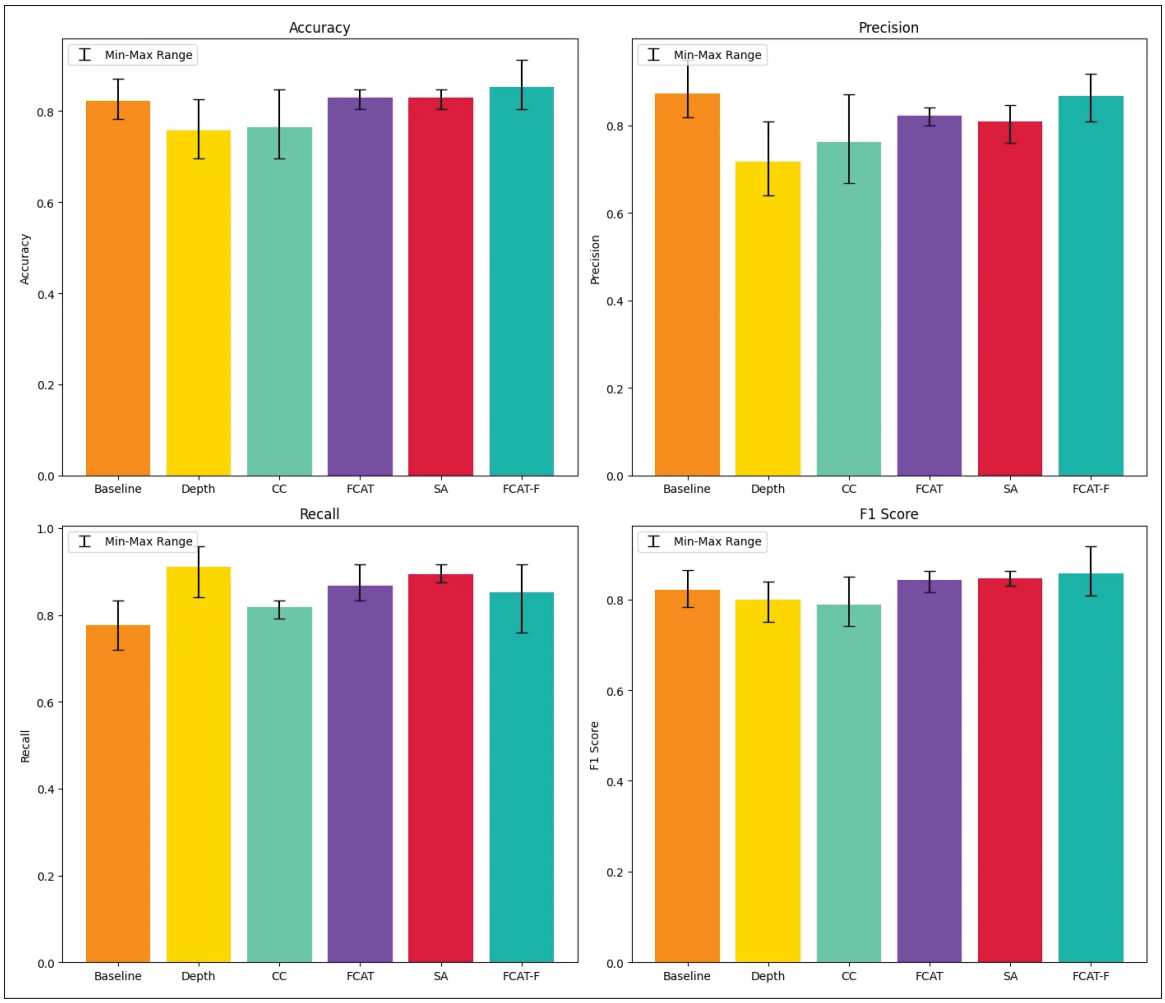

Figure 3.5 Performance comparison of ARD detection (X3D) system using different modality and different modality fusion schemes. The bars represent the average performance, with error bars indicating the min–max range of each metric across five folds

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.

3.6 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.

CHAPTER 4

IMPROVING ACUTE RESPIRATORY DISTRESS DETECTION IN PEDIATRIC ICUS USING AUTOMATED SPATIOTEMPORAL REGION EXTRACTION

Wajahat Nawaz[1] , Philippe Jouvet[2] , Rita Noumeir[1]

[1] Department of Electrical Engineering, École de Technologie Supérieure, 1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] Research Center at CHU Sainte-Justine Hospital, University of Montreal, 3175 Chem. de la Côte-Sainte-Catherine, Montréal, Québec, Canada H3T 1C5

Article Submitted in « *IEEE Journal of Biomedical and Health Informatics* » in October 2025

Abstract

Deep learning models for acute respiratory distress (ARD) detection in clinical videos face significant challenges in learning generalizable features due to imperfect medical datasets characterized by limited labeled data and spatiotemporal biases. These biases arise from visual artifacts (oxygen masks) and clinically irrelevant patient movements (repositioning, coughing) leading to spurious correlations and noisy labels. Prior studies show that ROI-based preprocessing enhances performance, but existing manual, axis-aligned methods are impractical, fail to mitigate spatial biases, and ignore temporal context. To address these limitations, we develop a fully automated spatiotemporal preprocessing pipeline that eliminates manual intervention while removing biases from clinical video data. The pipeline comprises two key components: (1) an oriented bounding box (OBB) detector for ROI localization with minimal background, and (2) a region-aware motion filter to exclude clips containing irrelevant movements. Our automated pipeline achieves high performance in both components: the OBB detector achieves localization with 84% mAP@50-95 (Mean Average Precision at IoU 0.5 to 0.95), and the motion component identifies irrelevant movements with 93% F_1score. For ARD detection, models using our preprocessing show substantial improvements; precision increases from 76% to 80% and specificity from 72% to 78% compared to axis-aligned. Class activation maps provide visual evidence that proposed preprocessing robustly redirects model attention from confounding artifacts to relevant regions. This study highlights the importance

of automated spatiotemporal preprocessing in guiding models to capture clinically relevant features instead of spurious artifacts, particularly under limited and imperfect labeled data. The fully automated pipeline further enables continuous patient monitoring in intensive care units, eliminating the need for manual intervention.

4.1 Introduction

Acute respiratory distress syndrome (ARDS) represents a critical medical emergency affecting approximately 200,000 people annually in the United States and 3 million worldwide. ARDS accounts for 10% of all intensive care unit (ICU) admissions and is responsible for at least 25% of cases requiring mechanical ventilation in hospital settings (Cleveland Clinic, 2025). In clinical practice, healthcare providers continuously monitor respiratory parameters; including breathing rate, chest wall movements, and other breathing patterns collectively referred to as acute respiratory distress (ARD) indicators.

Signs of chest retraction are critical visual indicators of ARD, particularly in neonates and young children. Medical staff typically rely on periodic visual assessment (often every 30 minutes), leading to inter-observer variability, delayed diagnosis, and subjectivity. The challenge is further exacerbated by high patient-to-nurse ratios in many intensive care units, particularly during pandemics and remote areas. This monitoring could be improved by automated ARD detection systems to ensure reliable diagnosis and continuous patient surveillance. Deep learning techniques are increasingly being employed to automate healthcare tasks, such as radiological imaging (Yahyatabar *et al.*, 2023), pathology (Araújo *et al.*, 2017), and physiological signs monitoring (Chavernac *et al.*, 2025), offering the potential to overcome many limitations of manual assessment approaches. This success has been largely attributed to the ability of these models to automatically learn hierarchical feature representations from large-scale datasets (Thian *et al.*, 2022). Large-scale datasets like ImageNet (Russakovsky *et al.*, 2015) have demonstrated strong generalization capabilities across diverse visual tasks.

However, obtaining large scale perfect labeled data in PICUs is challenging due to patient comfort constraints (Rehouma *et al.*, 2018), subjective annotation, and inter-observer variability. Beyond these general limitations, clinical video datasets introduce additional complexities specific to respiratory monitoring in PICUs. Cameras are typically mounted at a distance to minimize intrusion, resulting in a wide field of view that captures large portions of irrelevant background. Moreover, medical instruments and equipment introduce systematic biases between healthy and unhealthy patients. Oxygen masks, ventilators and other medical devices are predominantly present on patients with ARD, creating spurious correlations where models learn to associate equipment presence with positive ARD diagnosis rather than identifying genuine physiological indicators. Patients also exhibit clinically irrelevant movements, such as hand movements, crying episodes, or repositioning, which introduce temporal noise and mimic the clinical symptoms. These confounding factors significantly degrade datasets quality and introduce biases that prevent models from learning clinically relevant features. These challenges necessitate methods that explicitly handle the unique characteristics of PICU video data.

Transfer learning addresses limited labeled data by leveraging models pre-trained on large-scale datasets, showing effectiveness across medical imaging tasks (Araújo *et al.*, 2017; Vesal *et al.*, 2018; Cireşan *et al.*, 2013; Acharya & Basu, 2020; Lanjewar *et al.*, 2023; Sultani *et al.*, 2022). However, the effectiveness of TL can diminish significantly when there is a substantial semantic gap between the source and target domains. In clinical video settings, this gap is particularly pronounced. Video models pre-trained on action recognition datasets struggle to adapt on clinical videos datasets that contain medical equipment artifacts (biases), varying patient positions, and subtle respiratory motion. Furthermore, when applied to noisy clinical videos recordings, pre-trained models overfit to superficial patterns—such as the presence of oxygen masks and non-clinical movements features instead of learning the subtle chest wall movements. These domain mismatches limit generalization and highlight a critical need: tailored preprocessing techniques that guide models toward clinically relevant regions while suppressing noisy and irrelevant informations. It is especially helpful when we have limited amount of weakly-labeled datasets. A recent study (Nawaz *et al.*, 2024) has demonstrated that combining transfer learning

with appropriate data preprocessing improves model performance and prevents overfitting to irrelevant features.

Building on these insights, we propose an automated preprocessing pipeline that incorporates both spatial and temporal filtering strategies specifically designed for clinical respiratory monitoring. The spatial component uses an oriented bounding box (OBB) (Xia *et al.*, 2018) detector to automatically localize the thoracic-abdominal region, eliminating background clutter, irrelevant visual cues, and medical equipments that often confound model training. Unlike, a traditional axis-aligned bounding box that may include excessive background due to patient positioning, our OBB approach provides tighter, more accurate localization. On top of that, the temporal component employs region-aware motion filtering to exclude video segments containing excessive clinically irrelevant movements. By combining these spatial and temporal components, our preprocessing framework enables the model to focus on clinically meaningful features while filtering out irrelevant and noise information. This approach is particularly crucial in PICU settings where datasets are inherently biased toward medical instruments and manual preprocessing is infeasible. Our framework operates fully automatically from the detection of region of interest (ROI) to ARD.

4.1.1 Our Contributions

This work advances previous ARD detection research (Nawaz *et al.*, 2024; Nawaz, Albert, Jouvet & Noumeir, 2025) by introducing a fully automated spatial OBB-based ROI extraction that addresses fundamental challenges in clinical video analysis. Unlike (Dosso *et al.*, 2020), we propose a region-aware framework for detecting and filtering clinically irrelevant movements.

Our key contributions are:

1. We develop an automated ROI detection system for PICU setting that eliminates manual ROI selection while ensuring focus on clinically relevant regions.
2. We propose a region aware motion detection network to automatically detect video segments containing excessive clinically irrelevant movements.

3. We establish the quantitative and qualitative impact of spatiotemporal region extraction on ARD detection performance through extensive ablation studies, showing that clinical region isolation and clinically irrelevant movements clips removal are both necessary for reliable diagnosis.

The remainder of this paper is organized as follows: Section 4.2 reviews related work in transfer learning, ROI detection, and importance of motion filtering for medical video analysis. Section 4.3 presents our proposed automated preprocessing framework, detailing the OBB detection and region-aware motion filtering components. Section 4.4 describes the implementation details and experimental setup. Section 4.5 presents comprehensive experimental results and Discussion. Section 4.6 provides qualitative analysis through visualization studies. Finally, Section 4.8 concludes the paper and discusses future research directions.

4.2 Literature Review

We organize the related work into three categories that directly relate to our proposed framework: transfer learning for handling limited medical data, spatial preprocessing through ROI detection, and temporal preprocessing via motion filtering.

4.2.1 Transfer Learning

Transfer learning is widely adopted in vision tasks (Ren, He, Girshick & Sun, 2016; Chen, Papandreou, Kokkinos, Murphy & Yuille, 2017; Carreira & Zisserman, 2017) for faster convergence and improved performance in data-constrained settings. Transfer learning has become the de facto standard for addressing data scarcity in medical imaging. Early work by Bar et al. (Bar *et al.*, 2015) explored using off-the-shelf CNN features from ImageNet-trained models for chest pathology detection, combining them with handcrafted features like GIST descriptors. While this demonstrated the potential of pre-trained features, their approach achieved only marginal improvements over traditional methods, suggesting that direct feature transfer has limitations. Van Ginneken et al. (Van Ginneken, Setio, Jacobs & Ciompi, 2015) similarly applied

off-the-shelf CNN features for pulmonary nodule detection in CT scans, achieving 74% accuracy sufficient for screening but inadequate for clinical deployment.

Leveraging the fact that CNN final layers capture task-specific features while earlier layers learn general representations, researchers have employed fine-tuning techniques to adapt models for medical applications by replacing the final layers with task-specific layer. Shin et al. (Shin *et al.*, 2016) conducted a comprehensive comparison of three approaches: off-the-shelf features, training from scratch, and fine-tuning pre-trained models by changing the classification layer of those models. Their study on lymph node detection and interstitial lung disease classification revealed that fine-tuned models consistently outperformed both alternatives, achieving 85% accuracy compared to 78% for off-the-shelf features and 69% for random initialization. Recent study has extended transfer learning to video-based medical analysis. Nawaz et al. (Nawaz *et al.*, 2024) utilize 3D-CNNs models pre-trained on Kinetics-400 for ARD detection model training. Despite reducing overfitting, transfer learning remains susceptible to non-clinical features without proper preprocessing, as models exploit any dominant patterns regardless of clinical relevance.

4.2.2 ROI Detection

The concept of training deep learning models on ROI-cropped data rather than full frames has demonstrated significant benefits across various domains (Tu *et al.*, 2018; Tsou *et al.*, 2020; Nawaz *et al.*, 2024). Tu et al. (Tu *et al.*, 2018) demonstrated that training 3D-CNNs on human-centered ROIs substantially improved model performance compared to using full raw input. Extending this methodology to medical clinical tasks, Ben Salah et al. (Ben Salah, Jouvet & Noumeir, 2025) applied ROI-based training to remote photoplethysmography estimation, where models focused on specific facial areas (forehead and cheeks) delivered superior results compared to full-face training approaches. Most relevant to our work, Nawaz et al. (Nawaz *et al.*, 2024, 2025) employed ROI extraction as a preprocessing technique for ARD detection. Their findings (Nawaz *et al.*, 2024) revealed large improvements: models trained on thoracic-abdominal ROIs achieved 82% accuracy compared to 72% on cropped patient videos trainings 10% improvement attributed

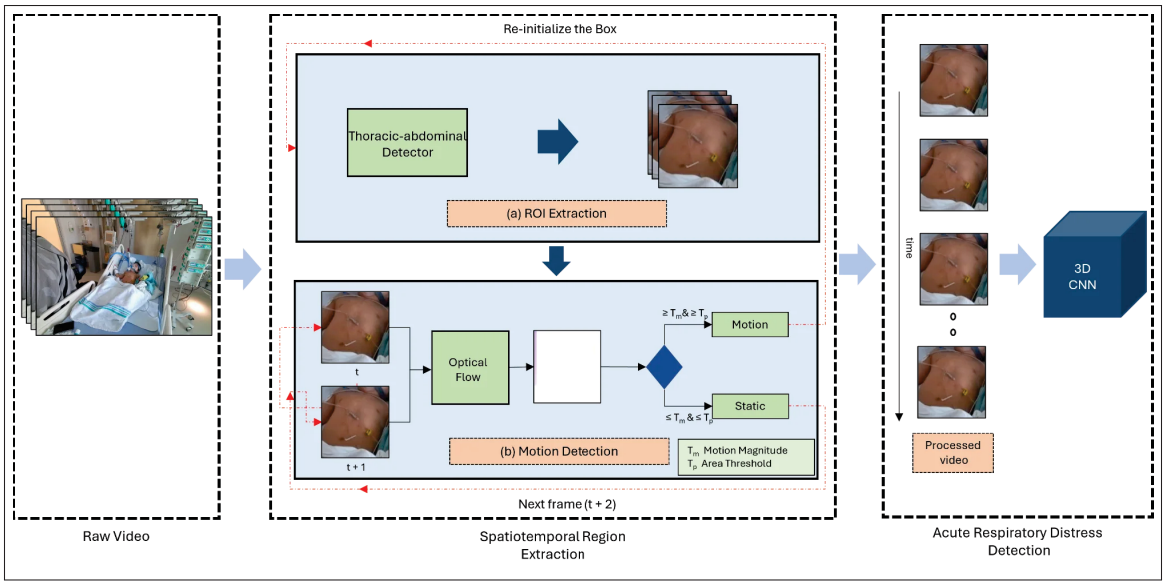

Figure 4.1 Overview of the proposed automated ARD detection system. The system comprises: (1) **Spatiotemporal Region Extraction** that performs OBB-based ROI detection for thoracic-abdominal localization and motion detection to detect non-clinical movements and ROI tracking, and (2) **ARD Detection** using 3D-CNNs classification. The framework analyzes optical flow within the detected ROI to classify frames as "Motion" or "Static" based on thresholds τ_m and τ_p. Excessive motion detection triggers automatic ROI re-initialization to maintain robust tracking. Static frames are accumulated into 6.4-second clips for ARD classification, enabling continuous automated monitoring

to eliminating background distractions. While ROI-based training demonstrably improves deep learning model performance. Existing approaches require manual selection or rely on pose estimation models that fail in clinical environments.

4.2.3 Motion Detection

Motion artifacts represent a major challenge in respiratory monitoring medical instruments, particularly for respiratory rate estimation where distinguishing between physiological and non-physiological movements is crucial. Hamilton et al. (Hamilton, Curley & Aimi, 2000) quantified the impact of motion artifacts on physiological parameter estimation, demonstrating that even minor patient movements can cause 40-60% error in heart rate estimation and 20-30% error in

respiratory rate measurement. Similarly, studies by Venkata et al. (Venkatasubramanian, Ranalli, Kirupaharan, Solanki & Mankodiya, 2024) demonstrate that non-clinical movements are one of the main causes of false alarms in NICU settings. Likewise, contactless respiratory monitoring devices are susceptible to motion artifacts, which degrade the quality of respiratory signals and compromise the accuracy of parameters such as respiratory rate and tidal volume (Massaroni, Nicolò, Sacchetti & Schena, 2021).

Various approaches have been developed to address motion artifacts in clinical monitoring. Hardware-based solutions, such as the pressure-sensitive mats proposed by Kyrollos et al. (Kyrollos, Greenwood, Harrold & Green, 2021), can differentiate between genuine physiological changes and motion artifacts with 91% accuracy. However, these approaches require additional equipment installation, calibration, and maintenance, limiting scalability across resource-constrained clinical settings.

Video-based motion detection offers a more scalable alternative. Dosso et al. (Dosso *et al.*, 2020) proposed a framework to reduce false alarms by detecting patient movements using RGB cameras and computer vision techniques. Their method analyzes frame-wide optical flow to identify when patients are moving. However, their approach treats all motion uniformly across the entire frame, a critical limitation for respiratory monitoring, where background motion can affect model performance. Existing motion detection either requires additional hardware, treats all motion uniformly, or lacks spatial awareness. Clinical respiratory monitoring needs to distinguish between respiratory motion (signal) and other movements (noise) within the same frame. Our region-aware approach analyzes motion exclusively within detected ROI boundaries, enabling this critical distinction.

4.3 Proposed Method

Our primary objective is to develop a robust system for the detection of ARD in PICU. We propose an automated spatiotemporal region extraction method that addresses the fundamental challenges of spatiotemporal biases inherent in clinical ARD datasets. Figure 4.1 illustrates the

complete pipeline of our methodology, which consists of two main sequential components: (1) a spatiotemporal region extraction framework and (2) an ARD detection system. The framework processes raw PICU video streams through sequential spatial and temporal filtering before classification. When excessive motion is detected, the system automatically reinitializes ROI detection to maintain accurate tracking during patient repositioning or clinical interventions. The detailed implementation of each component is presented in the following subsections.

4.3.1 Spatiotemporal Region Extraction

The spatiotemporal region extraction component performs automatic ROI selection, filters clinically irrelevant movements, and maintains robust tracking through adaptive re-initialization. Algorithm 4.1 details the automated spatiotemporal extraction preprocessing that transforms raw PICU camera streams into motion-filtered, ROI-focused video segments for ARD detection. The preprocessing operates through repeated cycles to acquire 6.4 second consecutive static clip segments suitable for ARD analysis. Initially, the OBB detector identifies the TA region in the incoming frames. Algorithm 4.1 (lines 8–14). Following ROI detection, the system collects frames while monitoring motion patterns—see Algorithm 4.1 (lines 20–39)—using optical flow. Optical flow computation occurs exclusively within the detected ROI boundaries and motion evaluation performed using the dual-threshold criteria defined in (4.8)–(4.9). Frames satisfying the static condition ($\mathcal{P}_{\text{moving}} < \tau_p$) are appended to $\mathcal{V}$. When three consecutive frames with excessive motion are detected, the system restarts the buffer $\mathcal{V}$ and initializes ROI re-detection (Algorithm 4.1, lines 33–35), resetting the collection process to handle patient repositioning. Upon accumulating 64 static frames (6.4 seconds at 10 FPS), the preprocessed segment is fed to the ARD detection network while the framework initiates a new collection cycle (lines 40–41). This continuous preprocessing approach maintains automated monitoring throughout the patient's PICU stay. The adaptive re-initialization mechanism handles patient movements and clinical interventions, eliminating manual preprocessing requirements while ensuring consistent input quality for ARD detection.

4.3.2 ROI Detection

The spatial preprocessing component focuses on automatically identifying and extracting the thoracic-abdominal region. We define the ROI as the anatomical region extending from the shoulders to the hip joints. To achieve precise localization while minimizing background inclusion, we employ the OBB approach (Xia *et al.*, 2018), which was originally developed for aerial image analysis and remote sensing applications, and has been used for thoracic-abdominal region extraction (Chavernac *et al.*, 2025). Unlike traditional axis-aligned bounding boxes that include substantial background regions when objects are rotated, OBB provides tighter region extraction. As demonstrated in Figure 4.2, when a rectangular object is rotated by 45°, the axis-aligned bounding box captures 2.45× more area compared to an OBB.

Specifically, for a rectangle with dimensions $w \times h$, the axis-aligned box must have sides of length $((w + h)/\sqrt{2})^2$ to fully contain the rotated object, resulting in an area of $((w + h)/\sqrt{2})^2$. In contrast, an OBB maintains the original area of $w \times h$ by rotating with the object. This excessive background inclusion increases the chance of visual noise that can compromise model performance.

OBB architecture consists of three components: (1) a backbone that generates multi-scale feature representations, (2) a feature pyramid network (FPN) that enhances feature maps by fusing multi-scale features, and (3) a detection head that predicts five parameters (x, y, w, h, θ) instead of the traditional four parameters. Here, x and y denote the coordinates of the bounding box center, w and h represent the width and height of the box, and θ is the rotation angle relative to the horizontal axis. The additional rotation parameter θ allows the box to tightly fit around patients regardless of their orientation on the bed.

The OBB detection employs a multi-component loss function that handles the additional rotation parameter. The total loss for training the OBB detection network is defined as:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{box} + \lambda_2 \mathcal{L}_{cls} + \lambda_3 \mathcal{L}_{obj} + \lambda_4 \mathcal{L}_{angle} \tag{4.1}$$

Algorithm 4.1 Spatiotemporal automated preprocessing pipeline

Input: Real-time RGB frame stream from camera
Output: Sequence of static ROI-cropped frames $\mathcal{V}$
$restart \leftarrow true$;
while $restart$ **do**
    Get frame F_{curr} from camera;
    while *ROI not detected* **do**
        Attempt to detect ROI (OBB) on F_{curr};
        if *ROI not detected* **then**
            Get next frame F_{curr} from camera;
        end if
    end while
    Initialize $static\_count \leftarrow 0$;
    Initialize $motion\_count \leftarrow 0$;
    Initialize $\mathcal{V} \leftarrow \emptyset$;
    $restart \leftarrow false$;
    while $static\_count < 64$ **do**
        Get next frame F_{next} from camera;
        Compute optical flow between ROI-cropped F_{curr} and F_{next};
        Compute flow magnitude and correct using (4.6), (4.7);
        Compute motion mask $\mathcal{M}_{\text{bin}}$ using (4.8);
        Calculate $\mathcal{P}_{moving}$ using (4.9);
        if $\mathcal{P}_{moving} < \tau_p$ **then**
            $static\_count \leftarrow static\_count + 1$;
            Append ROI-cropped F_{next} to $\mathcal{V}$;
            $motion\_count \leftarrow 0$;
        end if
        else
            $motion\_count \leftarrow motion\_count + 1$;
            if $motion\_count = 3$ **then**
                $restart \leftarrow true$;
                break;
            end if
        end if
        $F_{curr} \leftarrow F_{next}$;
    end while
    if $static\_count = 64$ **then**
        $restart \leftarrow true$;
    end if
end while
return $\mathcal{V}$;

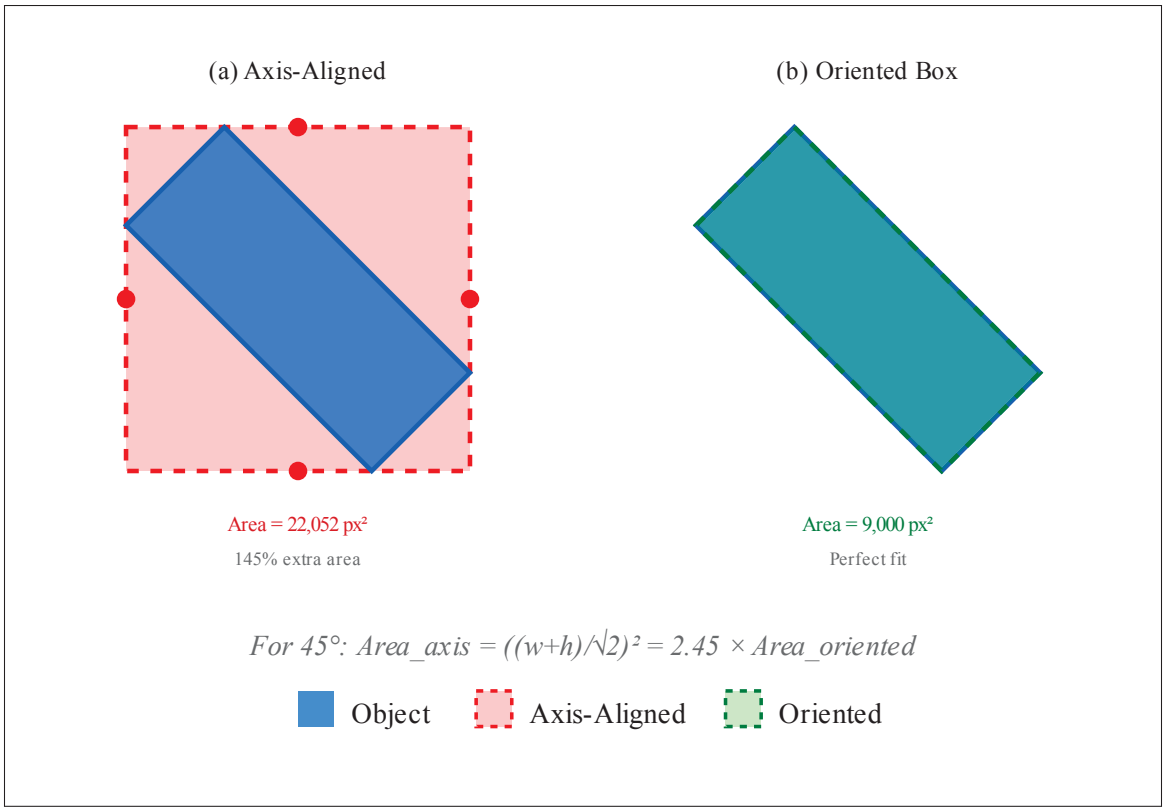

Figure 4.2 Comparison of axis-aligned versus OBB for a rectangle object rotated at 45°. (a) Axis-aligned box annotation in red. (b) Oriented box annotation in green.

Here, $\mathcal{L}_{box}$ denotes the bounding box regression loss for (x, y, w, h), $\mathcal{L}_{cls}$ is the classification loss, $\mathcal{L}_{obj}$ is the objectness loss, and $\mathcal{L}_{angle}$ is the rotation angle loss for θ. The coefficients λ_1, λ_2, λ_3, and λ_4 are weighting factors that control the relative contribution of each loss term during training.

The bounding box regression loss $\mathcal{L}_{box}$ combines spatial coordinate regression with rotated Intersection over Union (IoU):

$$\mathcal{L}_{box} = \mathcal{L}_{coord} + \mathcal{L}_{rIoU} \tag{4.2}$$

Where $\mathcal{L}_{coord}$ is the coordinate regression loss for (x, y, w, h) parameters, and $\mathcal{L}_{rIoU}$ is the rotated IoU loss that accounts for OBB overlap.

The angle loss $\mathcal{L}_{angle}$ handles the rotation parameter θ with periodic boundary considerations:

$$\mathcal{L}_{angle} = 1 - \cos(2(\theta_{pred} - \theta_{gt})) \tag{4.3}$$

Where θ_{pred} and θ_{gt} are the predicted and ground truth rotation angles, respectively. This formulation addresses angle periodicity where 0° and 180° represent identical orientations.

The objectness loss $\mathcal{L}_{obj}$ predicts the presence of an object in each candidate region using binary cross-entropy:

$$\mathcal{L}_{obj} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right], \tag{4.4}$$

where $y_i \in \{0, 1\}$ indicates the ground-truth object presence, p_i is the predicted probability, and N is the number of candidate regions. This loss ensures accurate distinction between object and background.

The classification loss $\mathcal{L}_{cls}$ predicts the object class for each detected region. For multi-class detection, it is computed using categorical cross-entropy:

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} \log(p_{i,c}), \tag{4.5}$$

where $y_{i,c} \in \{0, 1\}$ is the ground-truth label for class c, $p_{i,c}$ is the predicted probability for class c, C is the number of classes, and N is the number of detected objects. This loss ensures correct classification of detected objects.

4.3.3 Motion Detection

After spatially isolating the thoracic-abdominal region through ROI extraction, it is crucial to select temporally relevant video segments that exclude irrelevant patient movements, particularly

those not affecting the thoracic-abdominal region. For instance, if a patient slowly moves their hand, this motion does not impact the thoracic-abdominal region and should be filtered out. To achieve this objective, we analyze motion dynamics within the extracted ROI between consecutive video frames by computing dense optical flow vectors. Our region-aware approach analyzes motion exclusively within the detected ROI, distinguishing between clinically relevant thoracic-abdominal motion and irrelevant motion caused by sudden hand movements, repositioning, or other non-clinical activities.

We first compute the flow magnitude from dense optical flow using the Euclidean norm Eq. (4.6) of the horizontal and vertical flow components $F(x, y)$ as follow:

$$F(x, y) = \sqrt{u(x, y)^2 + v(x, y)^2} \tag{4.6}$$

where $u(x, y)$ and $v(x, y)$ represent the horizontal and vertical components of the optical flow vector $\mathbf{F}(x, y)$ at pixel (x, y).

To reduce camera motion artifacts, we then corrected flow magnitude by subtracting the global average flow magnitude using Eq. (4.7)

$$M(x, y) = \max\left(0,\ F(x, y) - \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F(i, j)\right) \tag{4.7}$$

where H and W are the height and width of the frame respectively, $N = H \times W$ is the total number of pixels in the frame, and $F(i, j)$ represents the flow magnitude at pixel coordinates (i, j). Then, we implement a dual-threshold strategy to identify clinically irrelevant movements. First, we determine which pixels within the ROI exhibit non-clinical motions by applying a motion magnitude threshold (4.8). Then we calculate the percentage of moving pixels (4.9).

$$\mathcal{M}_{\text{bin}}(x, y) = \begin{cases} 1, & \text{if } M(x, y) \geq \tau_m \quad \text{(large motion)} \\ 0, & \text{if } M(x, y) < \tau_m \quad \text{(small motion)} \end{cases} \tag{4.8}$$

$$\begin{aligned} \mathcal{P}_{\text{moving}} &= \frac{\text{Number of large motion pixels in ROI}}{\text{Total pixels in ROI}} \\ &= \frac{1}{|\text{ROI}|} \sum_{(x,y)\in\text{ROI}} \mathcal{M}_{\text{bin}}(x, y) \end{aligned} \tag{4.9}$$

where:

- τ_m is the motion magnitude threshold that separates small motion from large motion pixels
- $\mathcal{M}_{\text{bin}}(x, y)$ is a binary mask where 1 indicates a pixel has large motion (above threshold) and 0 indicates small motion (below threshold)
- |ROI| represents the total number of pixels within the region of interest
- $\mathcal{P}_{\text{moving}}$ is the percentage of pixels in the ROI that have large motion, ranging from 1 (all pixels moving) to 0 (all pixels static)

A frame is classified as static if $\mathcal{P}_{\text{moving}} \leq \tau_p$, where τ_p is the percentage threshold. The thresholds are determined through grid search on manually labeled datasets. This dual-threshold approach ensures robust selection of video segments containing minimal non-clinical motions. A motion event is considered valid only if three consecutive frames are classified as moving. A key advantage of our approach is that motion analysis is performed exclusively within the extracted thoracic-abdominal region, allowing the system to ignore irrelevant movements occurring outside this region. Consequently, movements such as hand gestures, foot movements, head adjustments, or caregiver activities in the background do not interfere with the motion assessment algorithms.

4.3.4 ARD Detection

Following spatial and temporal preprocessing, the filtered video segments undergo classification using our ARD detection network. We formulate ARD detection as a binary video classification problem and employ 3D Convolutional Neural Networks (3D-CNNs) due to their demonstrated

Table 4.1 Temporal and spatial dimension reduction in Channel-Separated CNN

Layer / Stage	**Output Size (T × H × W)**
Input	$T \times H \times W$
Conv1	$T \times \frac{H}{4} \times \frac{W}{4}$
ResBlock1	$T \times \frac{H}{4} \times \frac{W}{4}$
ResBlock2	$\frac{T}{2} \times \frac{H}{8} \times \frac{W}{8}$
ResBlock3	$\frac{T}{4} \times \frac{H}{16} \times \frac{W}{16}$
ResBlock4	$\frac{T}{8} \times \frac{H}{32} \times \frac{W}{32}$

capability in capturing spatiotemporal patterns essential for respiratory distress assessment. In contrast to conventional 3D-CNNs architectures such as SlowFast (Feichtenhofer *et al.*, 2019) and X3D (Xie *et al.*, 2018) which preserve full temporal resolution throughout the network, we adopt the channel-separated convolutional network (CSN) (Tran *et al.*, 2019) architecture.

Table 4.1 illustrates the systematic dimension reduction strategy employed by the CSN architecture. The network employs 3D convolutional filters that jointly process spatial and temporal dimensions while strategically incorporating temporal stride layers. Beginning with ResBlock2 as shown in Table 4.1, temporal resolution is progressively halved at each stage ($T \rightarrow T/2 \rightarrow T/4 \rightarrow T/8$), while spatial dimensions are initially downsampled by a factor of 4×, then progressively halved at each subsequent stage (skipping the ResBlock1). This progressive reduction enables the network to capture both fine-grained spatial features (chest wall movements within individual frames) and long temporal relation (breathing dynamics across sequences).

The resulting feature representations encode both spatial cues (visible chest retractions, breathing patterns) and temporal dynamics (respiratory rhythm irregularities, distress-related movements patterns) essential for accurate ARD classification. This architecture design is particularly suited for clinical respiratory monitoring, where subtle movements must be analyzed across extended time periods to distinguish abnormal breathing from normal breathing.

4.4 Implementation Details

This section details the experimental methodology and implementation procedures for each component of the proposed ARD detection system. All experiments were conducted using PyTorch on an NVIDIA A100 GPU with 32GB memory.

4.4.1 Dataset

We utilize ARD dataset collected at the Pediatric Intensive Care Unit of CHU Sainte-Justine Hospital, Montreal, Quebec, Canada (Boivin *et al.*, 2023). This study was approved by the Institutional Research Ethics Board of CHU Sainte-Justine (Ste-Justine REB number 2016-1242). Written informed parental consent was obtained for all participants prior to video recording. The complete dataset comprises 296 RGB-D video recordings (153 labeled, 143 unlabeled) from (PICU) admissions, each 30 seconds in duration. For this study, we only utilize the color (RGB) modality, as it provides sufficient visual information to capture respiratory distress signs.

1. **ROI Detection Dataset**: We compiled an OBB database of 1,272 images consisting of 400 external sourced images (Google searche using keywords "newborn in ICU", "newborn" and "kids") and 872 images extracted from the hospital's PICU video database. All images were manually annotated using LabelStudio (Tkachenko, Malyuk, Holmanyuk & Liubimov, 2020-2025) with OBB for two region categories: thoracic-abdominal region only, and region including facial area. Annotations were stored in eight-coordinate format (x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4) representing the four corners of each OBB.
2. **Motion Detection Dataset**:
   The 153 videos were segmented into 918 non-overlapping 5-second clips (6 clips per video). Clips were classified as either "moving" (containing patient body movements affecting the thoracic-abdominal region or clinical interventions) or "static" (minor movements not significantly impacting the thoracic-abdominal region). Manual labeling was performed by trained annotators.
3. **ARD Detection Dataset**:

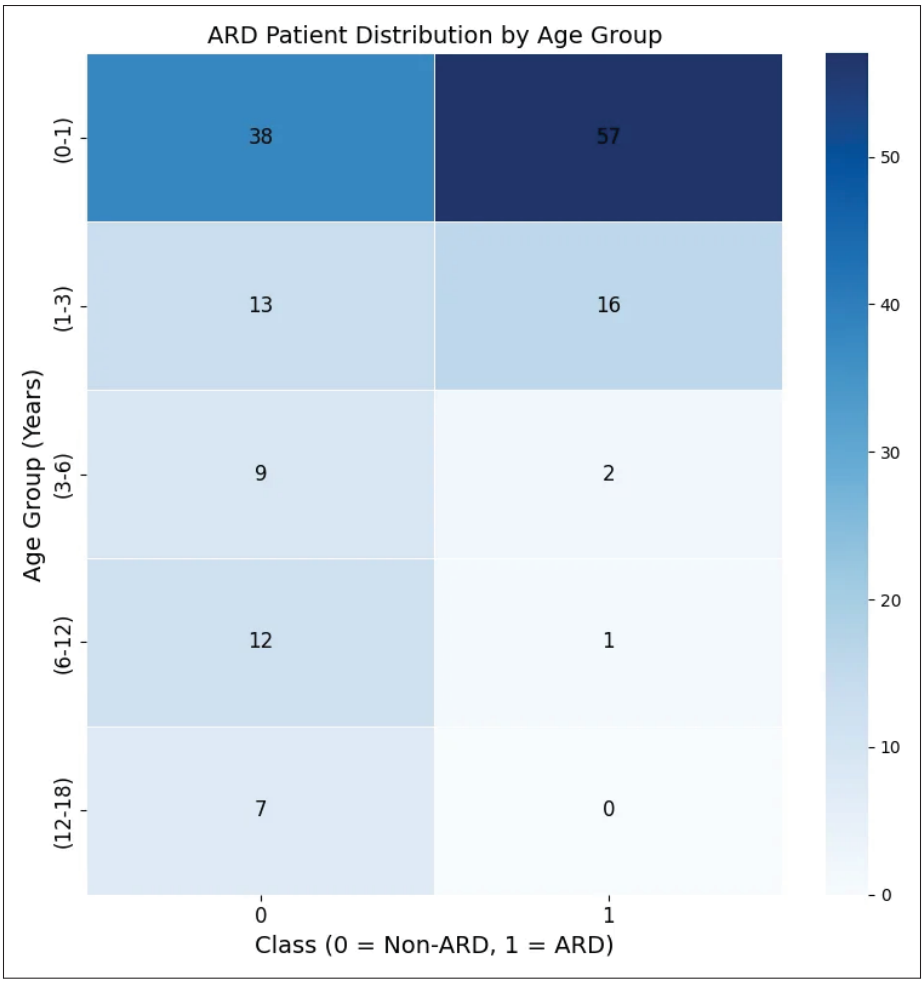

Figure 4.3 Distribution of Acute Respiratory Distress (ARD) patients across different age groups and clinical classes (0 = Non-ARD, 1 = ARD). The heatmap highlights the frequency of cases (darker blue shades indicating higher patient counts)

The 153 labeled videos include patients aged 0-18 years with ARD classification based on presence of chest retraction signs. Two experienced pediatric physicians independently reviewed each video for subcostal, intercostal, suprasternal, or substernal retractions. Videos with disagreement between reviewers were excluded to maintain annotation quality. The final distribution consists of 86 patients (56.2%) with chest retractions (ARD class) and 67 patients (43.8%) without retractions (non-ARD class). The class distribution by age group is illustrated Figure 4.3.

4.4.2 ROI Detection Implementation

We implemented OBB detection using YOLO11 variants (nano, small, medium) through the Ultralytics framework (Jocher & Qiu, 2024). The selection of YOLO11 architecture is motivated by its demonstrated efficiency in real-time object detection tasks. To enable OBB detection, we modified the detection head to predict an additional rotation parameter θ alongside the standard bounding box coordinates. OBB models' backbone are initialized using two different pre-training strategies to evaluate optimal feature transfer for clinical ROI detection: COCO Object Detection weights (standard object detection pre-training) and COCO Human Pose Estimation weights (keypoint detection pre-training focused on human body structure). We evaluated two distinct training configurations: single-class configuration training exclusively on thoracic-abdominal region detection (TA), and two-class configuration with joint training on both thoracic-abdominal and face-thoracic-abdominal regions (TA + FTA). The two-class approach enhances model feature discrimination capabilities by increasing task complexity. Data is partitioned on the patient level to prevent any date leakage. Training employed the multi-component OBB loss function (4.1)) with AdamW (Loshchilov & Hutter, 2017) optimizer (learning rate: 0.005, momentum: 0.937). Models underwent 100-epoch training at 1024×1024 resolution with batch size 16. Data augmentation included spatial scaling, rotation, brightness adjustment, and mosaic are employed to enhance robustness across varying clinical imaging conditions. Model performance was evaluated using mean average precision (mAP@50, mAP@50-95) with emphasis on mAP@50-95 for stringent localization accuracy assessment.

4.4.3 Motion Detection Implementation

Our motion detection employs RAFT-Flow (Teed & Deng, 2020) pretrained on FlyingChairs (Dosovitskiy *et al.*, 2015) and FlyingThings3D (Mayer *et al.*, 2016) datasets for optical flow computation. Moreover, we evaluate FlowFormer (Shi *et al.*, 2023) as an advanced optical flow algorithm for comparison. Motion detection is performed exclusively within the thoracic-abdominal region extracted by the ROI detector, as it is the primary region of interest where signs of chest retraction manifest. We evaluate the motion detection algorithm with three different

frame rates (10 FPS, 15 FPS, and 30 FPS) to determine optimal temporal resolution. The dual threshold parameters (motion magnitude threshold (τ_m) and percentage threshold (τ_p)) were optimized using a grid search approach with repeated random sub-sampling validation. The search space covered $\tau_m \in [0.10, 0.40]$ (0.05 increments) and $\tau_p \in [15\%, 50\%]$ (5% increments). For each repetition, the dataset was randomly partitioned into development (50%) and test (50%) sets. Optimal thresholds were determined on the development set by maximizing the F_1-score, and final performance metrics were computed on the held-out test set. We consider a patient as moving if motion is detected in three consecutive frames, which helps distinguish sustained non-clinical movements from transient motion artifacts.

4.4.4 ARD Detection Implementation

Our Acute Respiratory Distress detection system follows a similar data processing configuration as described in (Nawaz *et al.*, 2024), with the exception of utilizing OBB ROI extraction and motion filtering modules. The dataset is partitioned using a 70-30 patient-wise split for training and testing, stratified by age group and class to ensure equal distribution of patients across different age groups and classes. Videos are temporally subsampled at 10 FPS and divided into 14 overlapping clips to capture temporal dynamics. Additionally, videos are spatially resized to 256×256 with zero padding to maintain aspect ratio.

For motion-filtered datasets, we apply the motion detection framework to both training and test sets to generate static datasets by excluding clips containing clinically irrelevant movements. Patients with fewer than five overlapping clips are excluded to ensure sufficient temporal coverage. We fine-tune a channel-separated convolutional network using the AdamW (Loshchilov & Hutter, 2017) optimizer with a learning rate of 0.00001 and weight decay of 0.01. Training is conducted with a batch size of 128 using gradient accumulation. The model is trained for 30 epochs with binary cross-entropy loss and a label smoothing factor of 0.1. Data augmentation includes spatial transformations (rotation ±45°, random cropping 224×224, brightness adjustment ±0.3, scaling $1-1.25\times$, horizontal and vertical flipping) and temporal jittering to enhance model robustness.

Table 4.2 Performance evaluation of YOLO-v11-OBB models for ROI detection. Note:* Model is trained on one class

Model	Class	Box (P)	Box (R)	mAP@50	mAP@50-95
COCO 2017 Object Detection Pretrained Model					
YOLO-v11n*	TA	0.992	0.966	0.994	0.803
YOLO-v11n	TA	0.970	0.965	0.989	0.810
YOLO-v11s	TA	0.982	0.978	0.994	0.822
YOLO-v11m	TA	0.971	0.971	0.991	0.816
COCO Key Points 2017 Pretrained Model					
YOLO-v11s	TA	0.996	0.998	0.995	0.836
YOLO-v11m	TA	**0.993**	**0.996**	**0.995**	**0.840**

We performed a controlled ablation study comparing three ROI cropping strategies: (1) face-thoracic-abdominal (FTA), (2) thoracic-abdominal (TA), and (3) thoracic-abdominal with OBB (TA-OBB) with and without motion filtering. To isolate the effect of cropping methodology, all experiments used identical model architecture, hyperparameters, and training protocols. Model performance is assessed using F_1-score, recall, precision, and true negative rate with patient-wise evaluation to ensure robustness across different patient populations. We compute the average score across all clips to obtain the final prediction for each patient.

4.5 Results & Discussion

This section presents comprehensive results for all components of our proposed automated preprocessing framework for acute respiratory distress detection. We evaluate each component individually, followed by an analysis of the overall ARD system performance through comparative evaluation and ablation studies.

4.5.1 ROI Detection

Table 4.2 presents the quantitative evaluation of OBB detection for ROI localization across three experimental dimensions: training complexity (single vs. multi-class), model architecture scaling, and pre-training initialization strategies. Performance metrics include precision (P), recall (R), and mean average precision at IoU thresholds of 0.5 (mAP@50) and 0.5:0.95 (mAP@50-95).

The first two rows compare model performance trained on single-class versus two-class detection using YOLO-v11n with COCO object detection initialization, where the single-class model detects only the thoracic-abdominal region, while the two-class model detects both thoracic-abdominal and face-thoracic-abdominal regions. Multi-class training marginally improves mAP@50-95 from 0.803 to 0.810, while maintaining high precision (0.970) and recall (0.965). This improvement, though modest, indicates that distinguishing between thoracic-abdominal and face-thoracic-abdominal regions enhances the model's spatial understanding. The multi-task learning framework acts as an implicit regularizer, forcing the network to learn discriminative boundaries between anatomically adjacent regions rather than relying on coarse patient localization. Rows 2-4 evaluate YOLO-v11 variants (nano, small, medium) under identical two-class training conditions. YOLO-v11s achieves the best balance with mAP@50-95 of 0.822, outperforming both the lightweight nano (0.810) and the deeper medium variant (0.816). This non-monotonic scaling behavior suggests that ROI detection in clinical videos benefits from moderate model capacity; sufficient to capture anatomical boundaries without overfitting to dataset-specific artifacts. The small variant's superior performance (precision: 0.982, recall: 0.978) demonstrates that precise ROI localization does not require excessive parameterization. The most significant performance gains emerge from keypoint-based initialization (bottom rows). Models pre-trained on COCO keypoints consistently outperform object detection initialization: YOLO-v11s improves from mAP@50-95 of 0.822 to 0.836 (1.7% gain), while YOLO-v11m shows even greater improvement from 0.816 to 0.840 (2.9% gain). This superior performance stems from task alignment; keypoint detection networks learn to precisely localize anatomical landmarks (shoulders, chest, hips) that naturally delineate the thoracic-abdominal boundary.

Table 4.3 Performance evaluation of motion detection component

Model	**Frame Rate**	**Threshold**		**Evaluation Metrics**			
		τ_m	τ_p (%)	Accuracy	Recall	Precision	F_1 Score
RAFT Flow	10	0.40	20	0.962	0.944	0.903	0.923
	15	0.35	15	**0.964**	0.943	**0.913**	**0.928**
	30	0.25	15	0.946	0.865	0.905	0.884
Flow Former	10	0.35	20	0.959	**0.955**	0.885	0.919
	15	0.30	15	0.956	0.922	0.911	0.916
	30	0.25	15	0.934	0.831	0.8915	0.860

The keypoint-initialized YOLO-v11m achieves the highest overall performance (mAP@50-95: 0.840, precision: 0.993, recall: 0.996), demonstrating near-perfect ROI localization capability.

The consistent mAP@50 scores above 0.99 across all configurations indicate robust detection at standard IoU thresholds, essential for clinical reliability. The high mAP@50-95 scores ($\geq$ 80%) confirm precise boundary alignment crucial for eliminating background artifacts. Based on these results, we select YOLO-v11m with keypoint initialization for deployment, achieving 84% mAP@50-95 sufficient precision for automated clinical video preprocessing without manual intervention.

4.5.2 Motion Detection

Table 4.3 presents the performance of binary motion classification on 5-second video clips using the proposed motion detection approach. We employed RAFT-Flow (Teed & Deng, 2020) and FlowFormer (Shi *et al.*, 2023) for optical flow computation. Each segment is classified as either: (1) containing clinically relevant thoracic-abdominal motion suitable for respiratory analysis, or (2) containing irrelevant motion (e.g., patient repositioning, hand movements) that should be excluded. RAFT-Flow achieved optimal classification performance at 15 FPS, with accuracy of 0.964, precision of 0.913, and F_1-score of 0.928. This frame rate effectively captures

Table 4.4 Performance evaluation of ARD detection models with different preprocessing configurations. Confusion matrices are formatted as $\left[\begin{smallmatrix} TN & FP \\ FN & TP \end{smallmatrix}\right]$ where rows represent actual classes (0: non-ARD, 1: ARD) and columns represent predicted classes. TN: true negatives, FP: false positives, FN: false negatives, TP: true positives, TNR: true negative rate

Train → Test	F_1 **Score**	**Precision**	**Recall**	**TNR**	**Confusion Matrix**
Static-FTA → Static-FTA	0.800	0.762	0.842	0.722	$\left[\begin{smallmatrix} 13 & 5 \\ 3 & 16 \end{smallmatrix}\right]$
Static-TA → Static-TA	0.790	0.790	0.790	**0.778**	$\left[\begin{smallmatrix} 14 & 4 \\ 4 & 15 \end{smallmatrix}\right]$
Static-TA-OBB → Static-TA-OBB	0.821	0.800	0.842	**0.778**	$\left[\begin{smallmatrix} 14 & 4 \\ 3 & 16 \end{smallmatrix}\right]$
Moving-TA → Moving-TA	0.842	0.800	0.889	0.684	$\left[\begin{smallmatrix} 13 & 6 \\ 3 & 24 \end{smallmatrix}\right]$
Moving-TA → Static-TA	0.800	0.762	0.842	0.722	$\left[\begin{smallmatrix} 13 & 5 \\ 3 & 16 \end{smallmatrix}\right]$

respiratory motion patterns while filtering out irrelevant movements. At 10 FPS, accuracy remained high (0.962) but precision dropped slightly to 0.891, indicating some irrelevant motion was misclassified as clinically relevant. Performance declined at 30 FPS (recall = 0.865), where the high temporal resolution captured excessive non-clinical motion, making it harder to distinguish relevant from irrelevant segments. FlowFormer achieved its highest recall (0.955) at 10 FPS, successfully identifying most segments with clinically relevant motion. However, lower precision (0.885) indicates this sensitivity came with more false positives segments with only irrelevant motion being incorrectly classified as containing clinical motion. At 15 FPS, FlowFormer maintained balanced performance (accuracy = 0.956, F_1-score = 0.916), though both methods struggled at 30 FPS.

The analysis reveals that RAFT-Flow excels in precision-oriented metrics, making it preferred for applications requiring stringent false positive control. Moderate frame rates (15 FPS) combined with appropriately calibrated motion thresholds (τ_m = 0.35, τ_p = 15%) yield optimal detection robustness for respiratory monitoring applications.

4.5.3 Acute Respiratory Distress Detection

Table 4.4 presents classification performance across different preprocessing configurations and patient movements conditions. We evaluate three ROI strategies: Face-Thoracic-Abdominal (FTA) including the full upper body (Nawaz *et al.*, 2024), Thoracic-Abdominal (TA) using axis-aligned boxes, and OBB thoracic-abdominal (OBB-TA) with precisely aligned respiratory regions. The first three rows compare these configurations using static patient data for both training and testing. The lower rows demonstrate the detrimental impact of including clinically irrelevant movements clips in training data through cross-testing experiments between static and moving (static+moving) datasets.

The quantitative results reveal significant performance variations across ROI configurations. Static-FTA achieves an F_1 score of 0.800 with precision of 0.762, indicating the model's reliance on facial features compromises specificity. Static-TA demonstrates more balanced performance with an F_1 score of 0.790 and equal precision-recall (0.790). Most notably, Static-OBB-TA achieves the highest F_1 score of 0.821, with precision of 0.800 and recall of 0.842, while maintaining a true negative rate (TNR) of 0.778. This 3.9% improvement in F_1 score over standard TA validates the effectiveness of OBB in eliminating spatial biases through precise ROI localization. The lower rows of Table 4.4 reveal critical insights about training with moving datasets. When models are trained on both static and moving clips combined (Moving-TA $\rightarrow$ Moving-TA), the F_1 score reaches 0.842 with high recall of 0.889. However, TNR drops substantially to 0.684. A 9.4% decrease compared to static-only training. This degradation occurs because the combined dataset introduces movements biases, creating a spurious correlation between movements and positive diagnosis. The row shows the TNR rate improves when we exclude the movements clips from the testing data. Testing this moving trained model on static-only data (Moving-TA $\rightarrow$ static-TA)) partially recovers TNR to 0.722, confirming that movements triggers misclassifications.

4.6 Qualitative Analysis

We conducted a comprehensive qualitative comparison of class activation maps (CAMs) across three training configurations: (1) Face and Thoracic-Abdominal (FTA), which includes the full upper body region; (2) standard Thoracic-Abdominal (TA) using axis-aligned bounding boxes; and (3) OBB Thoracic-Abdominal (OBB-TA) with precisely aligned respiratory regions. Figure 4.4 illustrates representative cases demonstrating progressive improvement in feature localization, with each subfigure (a)–(c) displaying CAMs FTA (top), TA (middle), and OBB-TA (bottom). The qualitative analysis reveals critical differences in learned feature representations. Models trained on FTA regions consistently focus on spurious visual cues, particularly oxygen masks ubiquitous in PICU settings. In Figures 4.4(a) and (b), the FTA model (top row) exhibits strong activation patterns concentrated on mask regions while demonstrating minimal attention to the TA area where genuine respiratory distress manifests. This pattern indicates the model has incorrectly learned to associate medical equipment with ARD rather than identifying actual physiological indicators. Figure 4.4(c) presents a particularly revealing case: a patient wearing an oxygen mask but exhibiting no respiratory distress. The FTA model generates strong activations on the mask region, resulting in false positive classification. This misclassification exemplifies the fundamental limitation of including facial regions during training where the model develops spurious correlations with equipment presence rather than learning respiratory patterns. The TA configuration (middle row) demonstrates improved spatial focus compared to FTA. Activation maps shift from facial regions toward the TA area, though residual activations persist in clinically irrelevant regions including bedding, clothing folds, and peripheral medical equipment. These scattered activations highlight the inherent limitation of axis-aligned bounding boxes, which cannot precisely isolate the respiratory region due to geometric constraints and patients position.

Examining the sequential frames in each case reveals that OBB-TA maintains the most temporally consistent activation patterns. While FTA activations fluctuate based on mask visibility and TA shows variable attention across frames, OBB-TA demonstrates stable focus on the respiratory region throughout the breathing cycle. This temporal stability is crucial for reliable respiratory pattern analysis. The OBB-TA configuration (bottom row) achieves optimal spatial precision.

CAMs exhibit tight alignment with anatomical boundaries of respiratory motion, producing concentrated activation patterns specifically over the chest wall and upper abdomen where breathing-related movements originate. Unlike axis-aligned approaches, OBB-TA eliminates extraneous background regions, forcing the model to learn exclusively from respiratory regions.

4.7 Discussion

This study addresses fundamental challenges in developing deep learning models for clinical video analysis collected under strict ethical approval, which results in imperfect data for model training. Our quantitative and qualitative evaluations demonstrate that spatiotemporal preprocessing is essential and effectively mitigates biases present in imperfect medical datasets, enabling models to learn clinically relevant features rather than spurious correlations.

Prior studies have employed either manual data preprocessing or deep learning-based pose detectors for ROI-based preprocessing. However, these approaches are often impractical and tend to fail in clinical settings due to domain gaps (Chavernac *et al.*, 2025). Additionally, previous research has utilized controlled and manually filtered datasets, restricting deployment since the developed systems may fail in real-world settings. Our proposed automated spatiotemporal region extraction method addresses these problems by automatically preprocessing imperfect data for improved deep learning training. We implemented an OBB-based YOLO-11 object detector to spatially preprocess the dataset, trained on domain-specific data. Our OBB detector achieved 84% mAP@50-95 with keypoint-initialized pretrained models, demonstrating near-perfect precision (0.993) and recall (0.996). This ensures consistently high-quality input for subsequent processing stages, which is crucial because even minor background inclusion can introduce confounding artifacts from medical equipment and bedding patterns. In contrast to previous research that trained YOLO-OBB models exclusively on adult patients (Chavernac *et al.*, 2025) in controlled environments, we trained our model on patients ranging from 0 to 18 years old using real PICU datasets, enhancing applicability across pediatric age groups. To remove temporal noise, we implemented an optical flow-based method for detecting clinically irrelevant motion, inspired by Dosso et al. (Dosso *et al.*, 2020) using RGB cameras, as opposed to hardware-based approaches

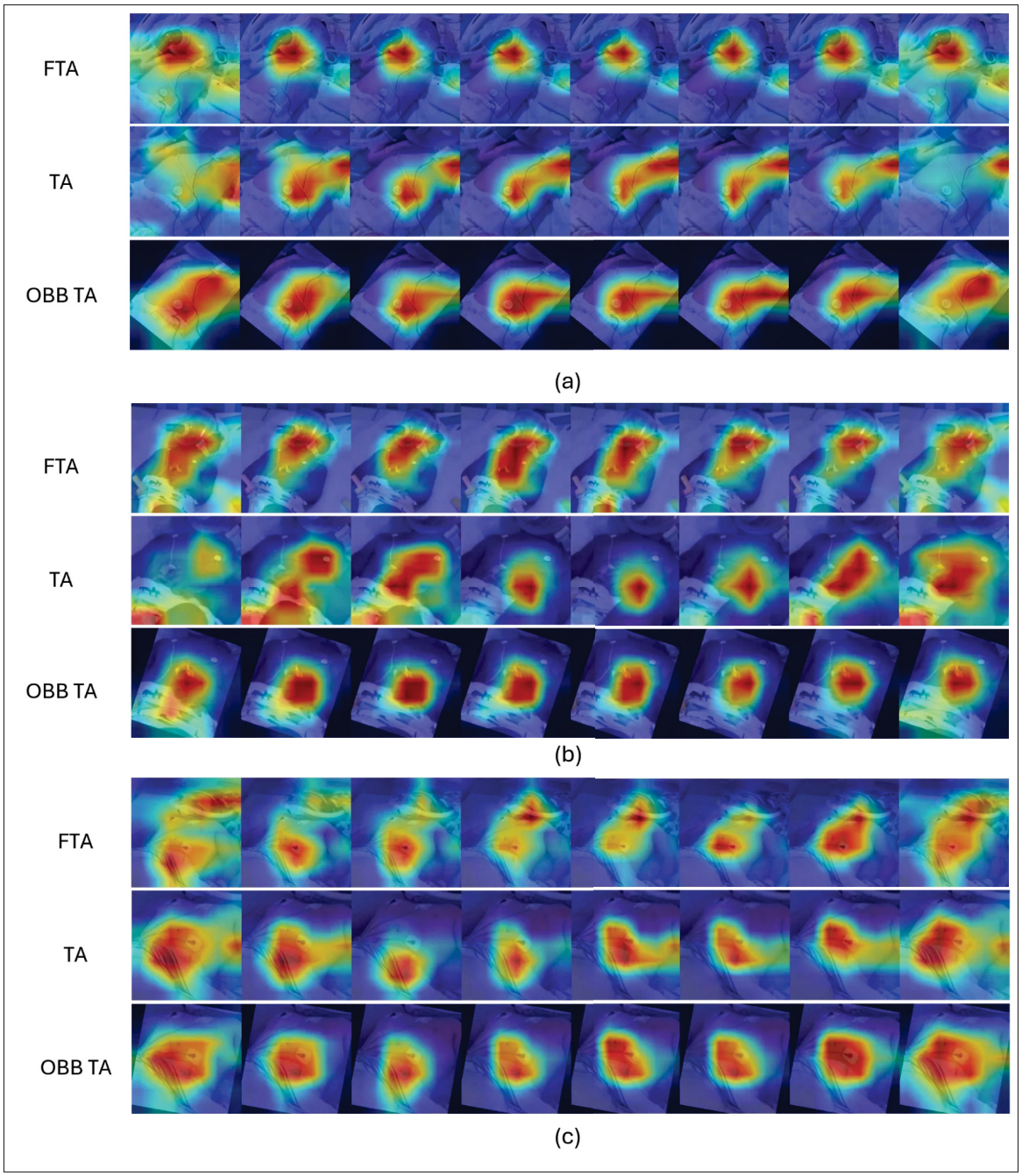

Figure 4.4 Class activation map (CAM) analysis comparing different training configurations for ARD detection across three representative clinical cases (a-c). For each case, the vertical sequence displays: Full-frame model (FTA) (top row), standard model (TA) model (middle row), and Oriented Bounding Box (OBB-TA) model (bottom row) respectively.

such as pressure-sensitive mats (Kyrollos *et al.*, 2021). Unlike previous frame-wide approaches, our method employs region-aware motion filtering invariant to background movements. Our motion detection component achieved optimal performance at 15 FPS (F_1-score: 93%). This region-aware approach advances beyond frame-wide methods (Dosso *et al.*, 2020) by ignoring background activity while maintaining sensitivity to thoracic-abdominal movements.

Applying OBB-based spatiotemporal preprocessing to ARD classification, we observed systematic improvement through precise ROI localization. Precision improved from 76% (Static-FTA (Nawaz *et al.*, 2024)) to 80% (Static-OBB-TA) while maintaining recall at 84%. Further analysis revealed that irrelevant movements are more common in ARD patients, creating a dataset bias where such movements become spuriously correlated with positive diagnosis due to limited data. Therefore, removing temporal biases is necessary for robust model training and testing. Critically, cross-testing experiments revealed that models trained on combined static and moving data with temporal noise (Moving-TA $\rightarrow$ Moving-TA: TNR = 0.684) suffered 9.4% true negative rate degradation compared to static-only training (Static-OBB-TA: TNR = 0.778). When tested exclusively on static data (Moving-TA $\rightarrow$ Static-TA: TNR = 0.722), specificity partially recovered, confirming that irrelevant movements create spurious correlations, causing models to associate movements with positive diagnosis rather than learning genuine respiratory indicators.

The CAM analysis Figure 4.4 provides visual evidence that models trained on FTA regions focus on oxygen masks rather than respiratory motion, even for non-ARD patients wearing masks. In contrast, OBB-TA configurations produce concentrated activations over chest wall and abdominal regions with temporally consistent patterns throughout breathing cycles. This validates that precise spatial preprocessing physically constrains the input space, forcing models to learn from clinically relevant regions while suppressing confounding artifacts. The fully automated spatiotemporal region extraction framework enables continuous PICU monitoring with the help of the ARD detection system, requiring no manual intervention and addressing workforce challenges in high patient-to-nurse ratio settings. Furthermore, the region-aware

motion detection component can also serve as context to detect false alarms generated by medical devices (Venkatasubramanian *et al.*, 2024).

4.8 Conclusion & Future Work

This paper presents a fully automated spatiotemporal preprocessing framework that addresses key challenges in training deep learning models on imperfect clinical video data. The framework includes an OBB detector for precise ROI localization and a region-aware motion filter to eliminate clinically irrelevant movements. By removing both spatial and temporal biases without manual intervention, the proposed approach enables continuous PICU monitoring under real-world conditions and prevents models from learning spurious correlations caused by medical equipment artifacts and patient motion.

Future work will extend this framework to additional physiological monitoring tasks, including respiratory rate and tidal volume estimation, with validation across diverse patient populations and clinical settings. We also plan to incorporate multi-task learning, enabling the model to simultaneously perform ARD detection and clinically irrelevant motion classification. This will allow the network to explicitly differentiate between respiratory distress patterns and incidental patient movements, enhancing robustness and feature discrimination. Data and reproducible codes are available upon request from Prof. Philippe Jouvet, M.D., Ph.D.

CONCLUSION AND RECOMMENDATIONS

This thesis presents the first fully automated system for continuous detection of acute respiratory distress (ARD) in pediatric intensive care units (PICUs). The proposed system utilizes RGB-D camera technology combined with deep learning algorithms to analyze chest wall motion patterns, enabling objective real-time assessment of respiratory effort. This work addresses a critical clinical gap: current ARD evaluation in pediatric settings depends primarily on subjective visual observation by healthcare professionals, which is prone to variability and limited by intermittent monitoring. Through three progressive and interrelated studies, this research demonstrates the feasibility, diagnostic performance, and clinical utility of deep learning-based systems for automated monitoring of visual symptoms of acute respiratory distress in children.

5.1 Summary of Key Contributions

The initial study confirmed the feasibility of automated acute respiratory distress detection by analyzing temporal video data. The results revealed that restricting the input to the thoraco-abdominal region significantly improved model performance, with accuracy increasing from 72.5% to 81.2%. Furthermore, it was determined that 6.4-second video clips capture a complete respiratory cycle, an essential duration for identifying acute respiratory distress. A comprehensive benchmarking of multiple 3D convolutional neural network (3D-CNN) architectures identified the channel-separated network (CSN-R101) as achieving optimal performance, with 81.9% accuracy, 79.8% precision, 89.1% recall, and an F_1-score of 84.0%. Class activation map (CAM) analysis confirmed that trained models focused on clinically relevant thoracic regions, providing strong evidence that the system learned meaningful respiratory patterns for decision-making.

The second study advanced the framework through multi-modal RGB-D fusion strategies. It was observed that depth data alone were insufficient to detect ARD due to the lack of surface texture information critical for distress characterization. However, by strategically integrating RGB

and depth modalities, the system overcame individual limitations. Among late fusion methods, feature concatenation with frozen base models (FCAT-F) yielded the highest performance, attaining an F_1-score of 86%, along with accuracy, precision, and recall rates of 85%, 87%, and 85%, respectively. These findings demonstrated that multi-modal integration enhances robustness and reliability under challenging conditions such as varying illumination and skin tones, a key requirement for real-world clinical deployment.

The third study addressed key challenges in clinical deployment by introducing automated spatiotemporal preprocessing. An oriented bounding box (OBB) based region of interest (ROI) detection network achieved an mAP@50–95 of 84%, ensuring precise localization of thoracic-abdominal regions while minimizing interference from medical equipment artifacts. In addition, a region-aware irrelevant motion detector achieved an F_1-score of 93%, automatically excluding frames with non-respiratory patient activity such as repositioning or limb movement. These advancements yielded notable improvements in detection precision (from 76% to 80%) and specificity (from 72% to 78%), underscoring the importance of intelligent preprocessing for robust and clinically viable ARD detection systems.

5.2 Clinical Significance and Impact

The proposed ARD detection system directly addresses critical limitations of current visual assessment methods used in pediatric intensive care units (PICUs). Continuous automated monitoring removes the need for constant clinician presence at the bedside, which is particularly significant in resource-limited settings such as rural hospitals, developing regions, or facilities facing specialist shortages. By automating visual assessment, the system reduces the risk of delayed recognition, a common cause of adverse outcomes under intermittent monitoring, thereby enabling earlier therapeutic intervention and potentially preventing the progression to respiratory failure.

A major strength of the system lies in its ability to minimize inter-observer variability, a well-documented issue in manual ARD assessment. Clinicians often rate respiratory distress severity subjectively, influenced by individual training and experience, and their judgment may be compromised under high workload or fatigue, which can lead to inconsistent evaluations. In contrast, the automated system standardizes decision criteria, ensuring uniform application across patients and clinical scenarios, thereby promoting consistency and reducing diagnostic ambiguity.

Another key contribution is the introduction of region-aware motion detection, which directly tackles one of the most significant practical challenges in continuous video-based monitoring: differentiating genuine respiratory distress from unrelated body movements or caregiver activity. In real PICU environments, patient repositioning and spontaneous movements can produce visual patterns resembling respiratory effort, contributing to alarm fatigue. The motion detection module contextualizes and filters these confounding activities, thereby excluding alarms triggered by non-respiratory movement. This selective attention reduces false positives, enhances reliability, and ensures that alerts correspond to clinically meaningful events.

Collectively, these innovations represent a significant step toward integrating objective, automated assessment tools into critical care workflows. The proposed system contributes to the growing field of AI-enabled respiratory monitoring, aligning with recent efforts that demonstrate AI's potential to enhance early ARD detection, improve diagnostic reliability, and support clinician decision-making.

5.3 Limitations and Challenges

Several important limitations require discussion. First, the study dataset consisted of 153 pediatric patients collected from a single institution (CHU Sainte-Justine Hospital), potentially limiting the generalizability of results across varied healthcare systems and demographic populations.

Although the dataset includes patients of diverse age and ethnic groups, its distribution is skewed toward younger patients, which may not fully capture the diversity seen across pediatric intensive care units globally. The dataset is relatively small and exhibits limited diversity, particularly with severe class imbalance skewed towards specific age groups. This imbalance creates substantial challenges when attempting to partition the data into three subsets (training, validation, and test) without critically compromising the representation of positive and negative examples. For instance, certain age ranges contain extremely sparse positive cases: only 3 positive patients fall within the 3-6 years age range compared to the negative cases, and merely 1 positive case exists in the 6-18 years age range, which is significantly less compared to 8 negative cases. Such extreme data scarcity makes traditional three-set (train-validation-test) splitting impractical, as it would either leave age groups entirely unrepresented in certain subsets or provide insufficient samples for robust model training and reliable performance evaluation. The repeated random sub-sampling validation approach provides a robust evaluation framework suitable for our dataset size. In summary, this study did not perform validation on a completely independent holdout cohort.

While the third study developed automated ROI detection achieving 84% mAP on the held-out dataset, this component still requires validation across diverse PICU layouts and camera configurations before practical deployment. Additionally, model performance is sensitive to the quality of the region of interest and motion detection component accuracy. Lastly, the research focused exclusively on detecting ARD presence or absence without quantifying distress severity. The binary classification approach, while clinically meaningful, does not provide clinicians with severity gradations that could guide therapeutic decisions. Extending the system to severity quantification would require expanded labeling efforts and model architectures supporting multi-class or regression-based predictions.

From a methodological standpoint, the study faces broader challenges shared across medical video deep learning research. Limited sample size constrains model generalization, as deep networks typically require extensive training data. Ethical and privacy constraints in pediatric medicine restrict large-scale and better quality data acquisition. To mitigate overfitting, this work adopted transfer learning, data augmentation, and region-focused pre-processing. However, these strategies can only partially offset inherent data scarcity. Weak labeling presents further constraints: available annotations represent video-level labels without frame-wise or respiratory cycle-level annotation, limiting the design of fine-grained temporal models. Moreover, despite interpretability tools such as class activation maps, deep learning models function largely as black boxes. In clinical practice, transparency is critical, as clinicians must understand and trust the model's rationale. Bridging this interpretability gap remains a major barrier to clinical adoption across all AI-driven diagnostic systems.

5.4 Future Directions

Several promising directions could further advance automated respiratory distress detection toward reliable clinical deployment.

Precise Respiratory Cycle Segmentation: Developing methods to automatically segment complete respiratory cycles from inspiration onset to expiration completion represents a critical next step. Cycle based segmentation enables models to analyze physiologically meaningful patterns rather than fixed-duration clips, improving both accuracy and computational efficiency. This targeted processing ensures that models learn discriminative temporal features of chest wall motion while reducing redundant data.

Cycle-Level Annotation: Future datasets should include respiratory cycle-level annotations by clinical experts. This would allow removal of ambiguous or transitional cycles, identification of intermittent distress episodes, and training of sequence models to track temporal evolution of

respiratory distress. Cycle specific annotation has the potential to enable early warning systems capable of detecting deterioration trends across time.

Retraction Sign Localization and Classification: Extending the system beyond binary classification toward spatially localizing and categorizing retraction patterns (subcostal, intercostal, substernal, suprasternal) would add direct clinical value. Temporal object detection or spatiotemporal segmentation frameworks could support this expansion, providing visual explanations through heatmaps or bounding boxes. Such functionality would enable severity quantification based on the number and type of retractions and enhance interpretability for clinical users by visually linking model decisions to physiological observations.

Computational Optimization and Deployment: Implementing model compression techniques, including knowledge distillation, quantization, and neural architecture search, will be crucial to ensure efficient operation in real-time and resource-limited settings without losing the accuracy of the system. Hardware optimization and edge inference pipelines could allow system deployment directly at the bedside without reliance on high performance servers, facilitating widespread adoption in low resource hospitals and developing regions.

This research demonstrates the feasibility and clinical promise of automated acute respiratory distress detection in pediatric intensive care through RGB-D imaging and deep learning. By addressing key limitations in current visual assessment methods, the proposed system contributes a critical step toward objective, continuous, and accessible respiratory monitoring. While challenges remain in data generalization, interpretability, and deployment, this work lays a strong foundation for future research toward real-time, AI-driven respiratory assessment capable of transforming pediatric critical care practice.

BIBLIOGRAPHY

Aarts, L. A., Jeanne, V., Cleary, J. P., Lieber, C., Nelson, J. S., Oetomo, S. B. & Verkruysse, W. (2013). Non-contact heart rate monitoring utilizing camera photoplethysmography in the neonatal intensive care unit—A pilot study. *Early human development*, 89(12), 943–948.

Acharya, J. & Basu, A. (2020). Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning. *IEEE transactions on biomedical circuits and systems*, 14(3), 535–544.

Addison, P. S., Smit, P., Jacquel, D., Addison, A. P., Miller, C. & Kimm, G. (2022). Continuous non-contact respiratory rate and tidal volume monitoring using a Depth Sensing Camera. *Journal of Clinical Monitoring and Computing*, 36(3), 657–665.

Addison, P. S., Antunes, A., Montgomery, D., Smit, P. & Borg, U. R. (2023). Robust non-contact monitoring of respiratory rate using a depth camera. *Journal of Clinical Monitoring and Computing*, 37(4), 1003–1010.

Adjabi, I., Ouahabi, A., Benzaoui, A. & Taleb-Ahmed, A. (2020). Past, present, and future of face recognition: A review. *Electronics*, 9(8), 1188.

Ahmad, S., Ullah, T., Ahmad, I., Al-Sharabi, A., Ullah, K., Khan, R. A., Rasheed, S., Ullah, I., Uddin, M. N. & Ali, M. S. (2022). A novel hybrid deep learning model for metastatic cancer detection. *Computational Intelligence and Neuroscience*, 2022(1), 8141530.

Ahmed, M. M., Oweidat, M., Okesanya, O. J., Alaswad, M., Abdelbar, S. M. M., Gill, P. & Alsabri, M. (2025). Barriers to Pediatric Emergency Care in Low-Resource Settings: A Narrative Review. *Sage Open Pediatrics*, 12, 30502225251336861.

Al-Naji, A. & Chahl, J. (2016). Remote respiratory monitoring system based on developing motion magnification technique. *Biomedical Signal Processing and Control*, 29, 1–10.

Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., Polónia, A. & Campilho, A. (2017). Classification of breast cancer histology images using Convolutional Neural Networks. *PLOS ONE*, 12(6), 1-14.

Bahrami, M. & Forouzanfar, M. (2022). Sleep apnea detection from single-lead ECG: A comprehensive analysis of machine learning and deep learning algorithms. *IEEE Transactions on Instrumentation and Measurement*, 71, 1–11.

Bai, Y.-W., Li, W.-T. & Chen, Y.-W. (2010). Design and implementation of an embedded monitor system for detection of a patient's breath by double webcams. *2010 IEEE International Workshop on Medical Measurements and Applications*, pp. 171–176.

Bar, Y., Diamant, I., Wolf, L., Lieberman, S., Konen, E. & Greenspan, H. (2015). Chest pathology detection using deep learning with non-medical training. *2015 IEEE 12th international symposium on biomedical imaging (ISBI)*, pp. 294–297.

Bartula, M., Tigges, T. & Muehlsteff, J. (2013). Camera-based system for contactless monitoring of respiration. *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2672–2675.

Behrman, R. E., Kliegman, R. M., Jenson, H. B. et al. (2000). *Nelson textbook of pediatrics*. WB Saunders Philadelphia.

Ben Salah, M. K., Jouvet, P. & Noumeir, R. (2025). rPPG Estimation: Vision Transformer With 3-D Temporal Central Difference. *IEEE Transactions on Instrumentation and Measurement*, 74, 1-13. doi: 10.1109/TIM.2025.3548236.

Benetazzo, F., Freddi, A., Monteriù, A. & Longhi, S. (2014). Respiratory rate detection algorithm based on RGB-D camera: theoretical background and experimental results. *Healthcare technology letters*, 1(3), 81–86.

Blonski, P. (2025). Pediatric ARDS Mortality Rates: The Role of Specialized Care. Retrieved from: https://ardsalliance.org/pediatric-ards-mortality-rates-the-role-of-specialized-care/.

Boivin, V., Shahriari, M., Faure, G., Mellul, S., Tiassou, E. D., Jouvet, P. & Noumeir, R. (2023). Multimodality video acquisition system for the assessment of vital distress in children. *Sensors*, 23(11), 5293.

Brieva, J., Ponce, H. & Moya-Albor, E. (2020). A contactless respiratory rate estimation method using a hermite magnification technique and convolutional neural networks. *Applied Sciences*, 10(2), 607.

Bubashait, M. & Hewahi, N. (2021). Urban Sound Classification Using DNN, CNN & LSTM a Comparative Approach. *2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, pp. 46–50.

Buolamwini, J. & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on fairness, accountability and transparency*, pp. 77–91.

Burghouts, G. J. & Schutte, K. (2013). Spatio-temporal layout of human actions for improved bag-of-words action detection. *Pattern Recognition Letters*, 34(15), 1861–1869.

Carreira, J. & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308.

Chavernac, F., Albert, K., Huy, H. V., Ramachandran, S., Noumeir, R. & Jouvet, P. (2025). Real-time current volume estimation system from an Azure Kinect camera in pediatric intensive care: Technical development. *Sensors*, 25(10), 3069.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4), 834–848.

Chen, M., Zhu, Q., Zhang, H., Wu, M. & Wang, Q. (2019). Respiratory rate estimation from face videos. *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp. 1–4.

Chen, X., Ma, W., Gao, W. & Fan, X. (2023). BAFNet: bottleneck attention based fusion network for sleep apnea detection. *IEEE Journal of Biomedical and Health Informatics*, 28(5), 2473–2484.

Cheng, J., Liu, R., Li, J., Song, R., Liu, Y. & Chen, X. (2023). Motion-robust respiratory rate estimation from camera videos via fusing pixel movement and pixel intensity information. *IEEE Transactions on Instrumentation and Measurement*, 72, 1–11.

Cireşan, D. C., Giusti, A., Gambardella, L. M. & Schmidhuber, J. (2013). Mitosis detection in breast cancer histology images with deep neural networks. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16*, pp. 411–418.

Cleveland Clinic. (2025). Acute Respiratory Distress Syndrome (ARDS). Retrieved from: https://my.clevelandclinic.org/health/diseases/15283-acute-respiratory-distress-syndrome-ards.

Das, S., Sharma, S., Dai, R., Bremond, F. & Thonnat, M. (2020). Vpn: Learning video-pose embedding for activities of daily living. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pp. 72–90.

Dey, R. & Salem, F. M. (2017). Gate-variants of gated recurrent unit (GRU) neural networks. *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pp. 1597–1600.

Di Tocco, J., Massaroni, C., Bravi, M., Miccinilli, S., Sterzi, S., Formica, D. & Schena, E. (2020). Evaluation of thoraco-abdominal asynchrony using conductive textiles. *2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pp. 1–5.

Diamond, M., Peniston, H. L., Sanghavi, D. K. & Mahapatra, S. (2025). Acute Respiratory Distress Syndrome. In StatPearls Editorial Team (Ed.), *StatPearls*. Treasure Island, FL: StatPearls Publishing. Retrieved from: https://www.ncbi.nlm.nih.gov/books/NBK436002/.

Diba, A., Fayyaz, M., Sharma, V., Karami, A. H., Arzani, M. M., Yousefzadeh, R. & Van Gool, L. (2017). Temporal 3d convnets: New architecture and transfer learning for video classification. *arXiv preprint arXiv:1711.08200*, 1-9.

Donoso, A., Arriagada, D., Contreras, D., Ulloa, D. & Neumann, M. (2016). Respiratory monitoring of pediatric patients in the Intensive Care Unit. *Boletín médico del Hospital Infantil de México*, 73(3), 149–165.

Dosovitskiy, A., Fischer, P., Ilg, E., Häusser, P., Hazırbaş, C., Golkov, V., v.d. Smagt, P., Cremers, D. & Brox, T. (2015). FlowNet: Learning Optical Flow with Convolutional Networks. *IEEE International Conference on Computer Vision (ICCV)*. Retrieved from: http://lmb.informatik.uni-freiburg.de/Publications/2015/DFIB15.

Dosso, Y. S., Aziz, S., Nizami, S., Greenwood, K., Harrold, J. & Green, J. R. (2020). Video-based neonatal motion detection. *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 6135–6138.

Douglas, I. S., Mehta, A. & Mansoori, J. (2024). Policy Proposals for Mitigating Intensive Care Unit Strain: Insights from the COVID-19 Pandemic. *Annals of the American Thoracic Society*, 21(12), 1633–1642.

Edwards, M. O., Kotecha, S. J. & Kotecha, S. (2013a). Respiratory distress of the term newborn infant. *Paediatric respiratory reviews*, 14(1), 29–37.

Edwards, M. O., Kotecha, S. J. & Kotecha, S. (2013b). Respiratory distress of the term newborn infant. *Paediatric respiratory reviews*, 14(1), 29–37.

Ernstmeyer, K. & Christman, E. [Open Resources for Nursing (Open RN)]. (2021). Nursing Skills. Retrieved from: https://www.ncbi.nlm.nih.gov/books/NBK593192/.

Farzaneh, N., Ansari, S., Lee, E., Ward, K. R. & Sjoding, M. W. (2023). Collaborative strategies for deploying artificial intelligence to complement physician diagnoses of acute respiratory distress syndrome. *NPJ Digital Medicine*, 6(1), 62.

Feichtenhofer, C. (2020). X3d: Expanding architectures for efficient video recognition. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 203–213.

Feichtenhofer, C., Pinz, A. & Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1933–1941.

Feichtenhofer, C., Fan, H., Malik, J. & He, K. (2019). Slowfast networks for video recognition. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6202–6211.

Feng, K., Qin, H., Wu, S., Pan, W. & Liu, G. (2020). A sleep apnea detection method based on unsupervised feature learning and single-lead electrocardiogram. *IEEE Transactions on Instrumentation and Measurement*, 70, 1–12.

Fiedler, M.-A., Rapczyński, M. & Al-Hamadi, A. (2020). Fusion-based approach for respiratory rate recognition from facial video images. *IEEE Access*, 8, 130036–130047.

Fiedler, M.-A., Werner, P., Rapczyński, M. & Al-Hamadi, A. (2023). Deep face segmentation for improved heart and respiratory rate estimation from videos. *Journal of Ambient Intelligence and Humanized Computing*, 14(7), 9383–9402.

Fonck, S., Fritsch, S., Nottenkämper, G. & Stollenwerk, A. (2023). Implementation of ResNet-50 for the Detection of ARDS in Chest X-Rays using transfer-learning. *Proceedings on automation in medical engineering*, 2(1), 742–742.

Gabruseva, T., Poplavskiy, D. & Kalinin, A. (2020). Deep learning for automatic pneumonia detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 350–351.

Gao, H., Wu, X., Geng, J. & Lv, Y. (2022). Remote heart rate estimation by signal quality attention network. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2122–2129.

Grooby, E., Sitaula, C., Tan, K., Zhou, L., King, A., Ramanathan, A., Malhotra, A., Dumont, G. A. & Marzbanrad, F. (2022). Prediction of neonatal respiratory distress in term babies at birth from digital stethoscope recorded chest sounds. *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 4996–4999.

Gupta, P., Bhowmick, B. & Pal, A. (2017). Accurate heart-rate estimation from face videos using quality-based fusion. *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 4132–4136.

Hamilton, P. S., Curley, M. & Aimi, R. (2000). Effect of adaptive motion-artifact reduction on QRS detection. *Biomedical instrumentation & technology*, 34(3), 197–202.

Hammer, J. (2013). Acute respiratory failure in children. *Paediatric respiratory reviews*, 14(2), 64–69.

Hart, A. & Lee, E. Y. (2019). Pediatric chest disorders: practical imaging approach to diagnosis. *Diseases of the Chest, Breast, Heart and Vessels 2019-2022: Diagnostic and Interventional Imaging*, 107–125.

Harte, J. M., Golby, C. K., Acosta, J., Nash, E. F., Kiraci, E., Williams, M. A., Arvanitis, T. N. & Naidu, B. (2016). Chest wall motion analysis in healthy volunteers and adults with cystic fibrosis using a novel Kinect-based motion tracking system. *Medical & biological engineering & computing*, 54(11), 1631–1640.

He, K., Gkioxari, G., Dollár, P. & Girshick, R. (2017). Mask r-cnn. *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969.

Health Recovery Solutions. (2025). Overcoming Healthcare Workforce Shortages through Remote Monitoring. Retrieved from: https://www.healthrecoverysolutions.com/blog/overcoming-healthcare-workforce-shortages-with-remote-monitoring.

Hedstrom, A. B., Gove, N. E., Mayock, D. E. & Batra, M. (2018). Performance of the Silverman Andersen Respiratory Severity Score in predicting PCO2 and respiratory support in newborns: a prospective cohort study. *Journal of Perinatology*, 38(5), 505–511.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735–1780.

Hon, K. L., Leung, K. K. Y., Oberender, F. & Leung, A. K. (2021). Paediatrics: how to manage acute respiratory distress syndrome. *Drugs in Context*, 10.

Hu, J.-F., Zheng, W.-S., Pan, J., Lai, J. & Zhang, J. (2018). Deep bilinear learning for rgb-d action recognition. *Proceedings of the European conference on computer vision (ECCV)*, pp. 335–351.

Huang, D., Zeng, Y., Zhu, Y., Song, X., Pan, L., Yang, J., Wang, Y., Lu, H. & Wang, W. (2024). Camera-Based Respiratory Imaging System for Monitoring Infant Thoracoabdominal Patterns of Respiration. *IEEE Journal of Biomedical and Health Informatics*, 1-14. doi: 10.1109/JBHI.2024.3482569.

Hurtado, D. E., Chavez, J. A., Mansilla, R., Lopez, R. & Abusleme, A. (2020). Respiratory volume monitoring: A machine-learning approach to the non-invasive prediction of tidal volume and minute ventilation. *IEEE Access*, 8, 227936–227944.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J. & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. *arXiv preprint arXiv:1602.07360*, 1-13.

Islam, M. M. & Iqbal, T. (2020). Hamlet: A hierarchical multimodal attention-based human activity recognition algorithm. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10285–10292.

Jaeger, V. K., Lebrecht, D., Nicholson, A. G., Wells, A., Bhayani, H., Gazdhar, A., Tamm, M., Venhoff, N., Geiser, T. & Walker, U. A. (2019). Mitochondrial DNA mutations and respiratory chain dysfunction in idiopathic and connective tissue disease-related lung fibrosis. *Scientific reports*, 9(1), 5500.

Jaimez, M., Kerl, C., Gonzalez-Jimenez, J. & Cremers, D. (2017). Fast odometry and scene flow from RGB-D cameras based on geometric clustering. *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3992–3999.

Ji, S., Xu, W., Yang, M. & Yu, K. (2012). 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1), 221–231.

Jocher, G. & Qiu, J. (2024). Ultralytics YOLO11 (Version 11.0.0). Retrieved from: https://github.com/ultralytics/ultralytics.

Joze, H. R. V., Shaban, A., Iuzzolino, M. L. & Koishida, K. (2020). MMTM: Multimodal transfer module for CNN fusion. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13289–13299.

Kapus, K., Rárosi, F., Novák, Z., Peták, F. & Tolnai, J. (2025). Monitoring respiratory function with telemedicine devices in asthmatic children. *Frontiers in Medicine*, 12, 1604909.

Kazhdan, M., Bolitho, M. & Hoppe, H. (2006). Poisson surface reconstruction. *Proceedings of the fourth Eurographics symposium on Geometry processing*, 7, 0.

Keskes, O. & Noumeir, R. (2021). Vision-based fall detection using st-gcn. *IEEE Access*, 9, 28224–28236.

Khalid, M. U. & Yu, J. (2018). Multi-Modal Three-Stream Network for Action Recognition. *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 3210-3215.

Khalid, N., Ghadi, Y. Y., Gochoo, M., Jalal, A. & Kim, K. (2021). Semantic recognition of human-object interactions via Gaussian-based elliptical modeling and pixel-level labeling. *IEEE Access*, 9, 111249–111266.

Kini, J., Fleischer, S., Dave, I. & Shah, M. (2023). Ensemble Modeling for Multimodal Visual Action Recognition. *arXiv preprint arXiv:2308.05430*, 1-7.

Klaser, A., Marszałek, M. & Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. *BMVC 2008-19th British Machine Vision Conference*, pp. 275–1.

Kumar, A. & Vishnu. (1996). Epidemiology of respiratory distress of newborns. *The Indian Journal of Pediatrics*, 63(1), 93–98.

Kyrollos, D. G., Greenwood, K., Harrold, J. & Green, J. R. (2021). Detection of false alarms in the NICU using pressure sensitive mat. *2021 IEEE sensors applications symposium (SAS)*, pp. 1–5.

Lanjewar, M. G., Panchbhai, K. G. & Charanarur, P. (2023). Lung cancer detection from CT scans using modified DenseNet with feature selection methods and ML classifiers. *Expert Systems with Applications*, 224, 119961.

Lee, Y. S., Pathirana, P. N., Evans, R. J. & Steinfort, C. L. (2015). Noncontact detection and analysis of respiratory function using microwave Doppler radar. *Journal of Sensors*, 2015(1), 548136.

Liu, G. & Guo, J. (2019). Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337, 325–338.

Liu, L., Lu, L., Luo, J., Zhang, J. & Chen, X. (2014). Enhanced Eulerian video magnification. *2014 7th International Congress on Image and Signal Processing*, pp. 50-54. doi: 10.1109/CISP.2014.7003748.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S. & Hu, H. (2022). Video swin transformer. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3202–3211.

Loshchilov, I. & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 1-19.

Mahbub, U., Imtiaz, H. & Ahad, M. A. R. (2011). An optical flow based approach for action recognition. *14th international conference on computer and information technology (ICCIT 2011)*, pp. 646–651.

Massaroni, C., Lo Presti, D., Formica, D., Silvestri, S. & Schena, E. (2019). Non-contact monitoring of breathing pattern and respiratory rate via RGB signal measurement. *Sensors*, 19(12), 2758.

Massaroni, C., Nicolò, A., Sacchetti, M. & Schena, E. (2021). Contactless Methods For Measuring Respiratory Rate: A Review. *IEEE Sensors Journal*, 21(11), 12821-12839.

Mateu-Mateus, M., Guede-Fernandez, F., Garcia-Gonzalez, M. A., Ramos-Castro, J. J. & Fernández-Chimeno, M. (2020). Camera-based method for respiratory rhythm extraction from a lateral perspective. *IEEE Access*, 8, 154924–154939.

Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A. & Brox, T. (2016). A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4040–4048.

Mazurowski, M. A., Buda, M., Saha, A. & Bashir, M. R. (2019). Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. *Journal of magnetic resonance imaging*, 49(4), 939–954.

McCollum, E. D. & Ginsburg, A. S. (2017). Outpatient management of children with World Health Organization chest indrawing pneumonia: implementation risks and proposed solutions. *Clinical Infectious Diseases*, 65(9), 1560–1564.

Merler, M., Ratha, N., Feris, R. S. & Smith, J. R. (2019). Diversity in faces. *arXiv preprint arXiv:1901.10436*, 1-29.

Milesky, J., Sharma, R., Rosen, M., Zhang, A. & Bass, E. B. (2025). Acute care nursing staff shortages that compromise patient-to-nurse ratios: A Making Healthcare Safer rapid response review. *Journal of Patient Safety and Risk Management*, 25160435251358976.

Mirabile, V. et al. (2023). Respiratory Failure. Retrieved from: https://www.ncbi.nlm.nih.gov/books/NBK526127/.

National Heart, Lung, and Blood Institute (NHLBI). (2025). Acute Respiratory Distress Syndrome (ARDS)-Symptoms and Causes. Retrieved from: https://www.nhlbi.nih.gov/health/ards/symptoms/.

Nawaz, W., Jouvet, P. & Noumeir, R. (2024). Automated Detection of Acute Respiratory Distress Using Temporal Visual Information. *IEEE Access*, 12(1), 142071-142082. doi: 10.1109/ACCESS.2024.3467266.

Nawaz, W., Albert, K., Jouvet, P. & Noumeir, R. (2025). Acute respiratory distress identification via multi-modality using deep learning. *Applied Sciences*, 15(3), 1512.

Nazir, S., Pateau, V., Bert, J., Clement, J.-F., Fayad, H., l'Her, E. & Visvikis, D. (2021). Surface imaging for real-time patient respiratory function assessment in intensive care. *Medical Physics*, 48(1), 142–155.

Ottaviani, V., Veneroni, C., Dellaca, R. L., Lavizzari, A., Mosca, F. & Zannin, E. (2022). Contactless monitoring of breathing pattern and thoracoabdominal asynchronies in preterm infants using depth cameras: A feasibility study. *IEEE Journal of Translational Engineering in Health and Medicine*, 10, 1–8.

Ouzar, Y., Djeldjli, D., Bousefsaf, F. & Maaoui, C. (2023). X-iPPGNet: A novel one stage deep learning architecture based on depthwise separable convolutions for video-based pulse rate estimation. *Computers in biology and medicine*, 154, 106592.

Pai, K.-C., Chao, W.-C., Huang, Y.-L., Sheu, R.-K., Chen, L.-C., Wang, M.-S., Lin, S.-H., Yu, Y.-Y., Wu, C.-L. & Chan, M.-C. (2022). Artificial intelligence–aided diagnosis model for acute respiratory distress syndrome combining clinical data and chest radiographs. *Digital Health*, 8, 20552076221120317.

Pardasani, R., Chaudhuri, R., Awasthi, N. & Goel, M. (2020). Machine Learning and Deep Learning Approaches to Quantify Respiratory Distress Severity and Predict Critical Alarms. *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 1–11.

Payán, D. R., Pedroza-Granados, J. & Domínguez-Cherit, G. [[Online; accessed 28-November-2025]]. (2007, Feb). Acute Respiratory Distress Syndrome and CT Scans. Retrieved on 2025-11-28 from: https://respiratory-therapy.com/disorders-diseases/critical-care/ards/acute-respiratory-distress-syndrome-and-ct-scans/.

Pilz, C. S., Zaunseder, S., Krajewski, J. & Blazek, V. (2018). Local group invariance for heart rate estimation from face videos in the wild. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 1254–1262.

Popescu, A.-C., Mocanu, I. & Cramariuc, B. (2020). Fusion mechanisms for human activity recognition using automated machine learning. *IEEE Access*, 8, 143996–144014.

Ramji, H. F., Hafiz, M., Altaq, H. H., Hussain, S. T. & Chaudry, F. (2023). Acute respiratory distress syndrome; a review of recent updates and a glance into the future. *Diagnostics*, 13(9), 1528.

Reamaroon, N., Sjoding, M. W., Gryak, J., Athey, B. D., Najarian, K. & Derksen, H. (2021). Automated detection of acute respiratory distress syndrome from chest X-Rays using Directionality Measure and deep learning features. *Computers in biology and medicine*, 134, 104463.

Rehouma, H., Noumeir, R., Bouachir, W., Jouvet, P. & Essouri, S. (2018). 3D imaging system for respiratory monitoring in pediatric intensive care environment. *Computerized Medical Imaging and Graphics*, 70, 17–28.

Rehouma, H., Noumeir, R., Essouri, S. & Jouvet, P. (2019a). Quantitative assessment of spontaneous breathing in children: evaluation of a depth camera system. *IEEE Transactions on Instrumentation and Measurement*, 69(7), 4955–4967.

Rehouma, H., Noumeir, R., Masson, G., Essouri, S. & Jouvet, P. (2019b). Visualizing and quantifying thoraco-abdominal asynchrony in children from motion point clouds: A pilot study. *IEEE Access*, 7, 163341–163357.

Ren, S., He, K., Girshick, R. & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6), 1137–1149.

Romano, C., Schena, E., Silvestri, S. & Massaroni, C. (2021). Non-contact respiratory monitoring using an RGB camera for real-world applications. *Sensors*, 21(15), 5126.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3), 211-252.

Sabokrou, M., Pourreza, M., Li, X., Fathy, M. & Zhao, G. (2021). Deep-hr: Fast heart rate estimation from face video under realistic conditions. *Expert Systems with Applications*, 186, 115596.

Saini, M. & Susan, S. (2022). Diabetic retinopathy screening using deep learning for multi-class imbalanced datasets. *Computers in Biology and Medicine*, 149, 105989.

Savran Kızıltepe, R., Gan, J. Q. & Escobar, J. J. (2021). A novel keyframe extraction method for video classification using deep neural networks. *Neural Computing and Applications*, 1–12.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J. & Napolitano, A. (2009). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE transactions on systems, man, and cybernetics-part A: systems and humans*, 40(1), 185–197.

Shen, Q., Qin, H., Wei, K. & Liu, G. (2021). Multiscale deep neural network for obstructive sleep apnea detection using RR interval from single-lead ECG signal. *IEEE Transactions on Instrumentation and Measurement*, 70, 1–13.

Shi, X., Huang, Z., Li, D., Zhang, M., Cheung, K. C., See, S., Qin, H., Dai, J. & Li, H. (2023). Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1599–1610.

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D. & Summers, R. M. (2016). Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging*, 35(5), 1285-1298.

Silverman, W. A. & Andersen, D. H. (1956). A controlled clinical trial of effects of water mist on obstructive respiratory signs, death rate and necropsy findings among premature infants. *Pediatrics*, 17(1), 1–10.

Simonyan, K. & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pp. 568–576.

Sjoding, M. W., Taylor, D., Motyka, J., Lee, E., Claar, D., McSparron, J. I., Ansari, S., Kerlin, M. P., Reilly, J. P., Shashaty, M. G. et al. (2021). Deep learning to detect acute respiratory distress syndrome on chest radiographs: a retrospective study with external validation. *The Lancet Digital Health*, 3(6), e340–e348.

Stanford Medicine Newborn Nursery. (2025). Lungs Chest Photo Gallery. Retrieved from: https://med.stanford.edu/newborns/professional-education/photo-gallery/lungs-chest.html.

Su, L., Wang, Y., Zhai, D., Shi, Y., Ding, Y., Gao, G., Li, Q., Yu, M. & Wu, H. (2024). Spatiotemporal Sensitive Network for Non-Contact Heart Rate Prediction from Facial Videos. *Applied Sciences*, 14(20), 9551.

Sultani, W., Nawaz, W., Javed, S., Danish, M. S., Saadia, A. & Ali, M. (2022). Towards Low-Cost and Efficient Malaria Detection. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20655–20664.

Szymański, P. & Kajdanowicz, T. (2017, 22 Sep). A Network Perspective on Stratification of Multi-Label Data. *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, 74(Proceedings of Machine Learning Research), 22–35.

Tarassenko, L., Villarroel, M., Guazzi, A., Jorge, J., Clifton, D. & Pugh, C. (2014). Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. *Physiological measurement*, 35(5), 807.

Taussig, L. M. & Landau, L. I. (2008). *Pediatric Respiratory Medicine E-Book*. Elsevier Health Sciences.

Teed, Z. & Deng, J. (2020). Raft: Recurrent all-pairs field transforms for optical flow. *European conference on computer vision*, pp. 402–419.

Thian, Y. L., Ng, D. W., Hallinan, J. T. P. D., Jagmohan, P., Sia, S. Y., Mohamed, J. S. A., Quek, S. T. & Feng, M. (2022). Effect of training data volume on performance of convolutional neural network pneumothorax classifiers. *Journal of Digital Imaging*, 35(4), 881–892.

Tkachenko, M., Malyuk, M., Holmanyuk, A. & Liubimov, N. [Open source software available from https://github.com/HumanSignal/label-studio]. (2020-2025). Label Studio: Data labeling software. Retrieved from: https://github.com/HumanSignal/label-studio.

Tran, D., Bourdev, L., Fergus, R. et al. (2015). Learning spatiotemporal features with 3d convolutional networks. *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y. & Paluri, M. (2018, June). A Closer Look at Spatiotemporal Convolutions for Action Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tran, D., Wang, H., Torresani, L. & Feiszli, M. (2019). Video classification with channel-separated convolutional networks. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5552–5561.

Transue, S., Nguyen, P., Vu, T. & Choi, M.-H. (2016). Real-time tidal volume estimation using iso-surface reconstruction. *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pp. 209–218.

Tsou, Y.-Y., Lee, Y.-A., Hsu, C.-T. & Chang, S.-H. (2020). Siamese-rPPG network: Remote photoplethysmography signal estimation from face videos. *Proceedings of the 35th annual ACM symposium on applied computing*, pp. 2066–2073.

Tu, Z., Xie, W., Qin, Q., Poppe, R., Veltkamp, R. C., Li, B. & Yuan, J. (2018). Multi-stream CNN: Learning representations based on human-related regions for action recognition. *Pattern Recognition*, 79, 32–43.

Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M. & Baik, S. W. (2017). Action recognition in video sequences using deep bi-directional LSTM with CNN features. *IEEE access*, 6, 1155–1166.

Van Ginneken, B., Setio, A. A., Jacobs, C. & Ciompi, F. (2015). Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. *2015 IEEE 12th International symposium on biomedical imaging (ISBI)*, pp. 286–289.

Vargas-Acevedo, C., Botero Marín, M., Jaime Trujillo, C., Hernández, L. J., Vanegas, M. N., Moreno, S. M., Rueda-Guevara, P., Baquero, O., Bonilla, C., Mesa, M. L. et al. (2024). Severity and mortality of acute respiratory failure in pediatrics: A prospective multicenter cohort in Bogotá, Colombia. *Health Science Reports*, 7(6), e1994.

Veeriah, V., Zhuang, N. & Qi, G.-J. (2015). Differential recurrent neural networks for action recognition. *Proceedings of the IEEE international conference on computer vision*, pp. 4041–4049.

Venkatasubramanian, K., Ranalli, T.-M., Kirupaharan, P., Solanki, D. & Mankodiya, K. (2024). Understanding the challenges nurses encounter with monitoring technologies in a NICU. *International Journal of Human–Computer Interaction*, 40(23), 8142–8165.

Vesal, S., Ravikumar, N., Davari, A., Ellmann, S. & Maier, A. (2018). Classification of breast cancer histology images using transfer learning. *Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings 15*, pp. 812–819.

Vitazkova, D., Foltan, E., Kosnacova, H., Micjan, M., Donoval, M., Kuzma, A., Kopani, M. & Vavrinsky, E. (2024). Advances in respiratory monitoring: a comprehensive review of wearable and remote technologies. *Biosensors*, 14(2), 90.

Wang, H. & Schmid, C. (2013). Action recognition with improved trajectories. *Proceedings of the IEEE international conference on computer vision*, pp. 3551–3558.

Wang, H., Kläser, A., Schmid, C. & Liu, C.-L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, 103(1), 60–79.

Wang, X., Hu, J.-F., Lai, J.-H., Zhang, J. & Zheng, W.-S. (2019). Progressive teacher-student learning for early action prediction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3556–3565.

Wang, Y., Acs, B., Robertson, S., Liu, B., Solorzano, L., Wählby, C., Hartman, J. & Rantalainen, M. (2022). Improved breast cancer histological grading using deep learning. *Annals of oncology*, 33(1), 89–98.

Warren, J. B. & Anderson, J. M. (2010). Newborn respiratory disorders. *Pediatrics in review*, 31(12), 487–496.

Weiyao, X., Muqing, W., Min, Z. & Ting, X. (2021). Fusion of skeleton and RGB features for RGB-D human action recognition. *IEEE sensors journal*, 21(17), 19157–19164.

Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M. & Zhang, L. (2018). DOTA: A large-scale dataset for object detection in aerial images. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3974–3983.

Xia, J. & Siochi, R. A. (2012). A real-time respiratory motion monitoring system using KINECT: proof of concept. *Medical physics*, 39(5), 2682–2685.

Xie, S., Sun, C., Huang, J., Tu, Z. & Murphy, K. (2018). Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. *Proceedings of the European conference on computer vision (ECCV)*, pp. 305–321.

Yahyatabar, M., Jouvet, P. & Cheriet, F. (2023). Joint classification and segmentation for an interpretable diagnosis of acute respiratory distress syndrome from chest x-rays. *Journal of Medical Imaging*, 10(5), 054504–054504.

Yang, Z. & Kurita, T. (2014). Improvements of Local Descriptor in HOG/SIFT by BOF Approach. *IEICE TRANSACTIONS on Information and Systems*, 97(5), 1293–1303.

Yehya, N. & Thomas, N. J. (2016). Relevant outcomes in pediatric acute respiratory distress syndrome studies. *Frontiers in pediatrics*, 4, 51.

Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R. & Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4694–4702.

Yuthong, A., Duangsoithong, R., Booranawong, A. & Chetpattananondh, K. (2019). Monitoring of volume of air in inhalation from Triflo using video processing. *IEEE Transactions on Instrumentation and Measurement*, 69(7), 4334–4347.

Zhang, J.-T., Tsoi, A.-C. & Lo, S.-L. (2014). Scale invariant feature transform flow trajectory approach with applications to human action recognition. *2014 International Joint Conference on Neural Networks (IJCNN)*, pp. 1197–1204.

Zhang, M. & Pang, M. (2023). Early prediction of acute respiratory distress syndrome complicated by acute pancreatitis based on four machine learning models. *Clinics*, 78, 100215.

Zhou, Q.-Y., Park, J. & Koltun, V. (2018). Open3D: A modern library for 3D data processing. *arXiv preprint arXiv:1801.09847*, 1-2.

Zhou, Y., Mei, S., Wang, J., Xu, Q., Zhang, Z., Qin, S., Feng, J., Li, C., Xing, S., Wang, W. et al. (2024). Development and validation of a deep learning-based framework for automated lung CT segmentation and acute respiratory distress syndrome prediction: a multicenter cohort study. *EClinicalMedicine*, 75.

Zhu, Z., Yu, L., Wu, W., Yu, R., Zhang, D. & Wang, L. (2022). MuRCL: Multi-instance reinforcement contrastive learning for whole slide image classification. *IEEE Transactions on Medical Imaging*, 42(5), 1337–1348.