

DM

Deep Learning-Based Estimation of Respiration from PPG

From Rate Extraction to Apnea Phenotype Detection in Wearable Devices

MASTER DISSERTATION

Sandro Canada Rodrigues
MASTER IN INFORMATICS ENGINEERING



UNIVERSIDADE da MADEIRA

A Nossa Universidade
www.uma.pt

July | 2026

**Deep Learning-Based Estimation
of Respiration from PPG**
From Rate Extraction to Apnea Phenotype Detection
in Wearable Devices

MASTER DISSERTATION

Sandro Canada Rodrigues

MASTER IN INFORMATICS ENGINEERING

SUPERVISOR

Fábio Rúben Silva Mendonça

CO-SUPERVISOR

Jeffrey Scott Durmer



FACULDADE DE CIÊNCIAS EXATAS E DA ENGENHARIA

MASTER OF SCIENCE DEGREE IN INFORMATICS ENGINEERING

**Deep Learning-Based Estimation of Respiration
from PPG: From Rate Extraction to Apnea
Phenotype Detection in Wearable Devices**

Sandro Canada Rodrigues

Supervised by:

Prof. Doutor Fábio Rúben Silva Mendonça (Supervisor)

Dr. Jeffrey Scott Durmer (Co-supervisor)

Constituição do júri de provas públicas:

Karolina Baras, Presidente

Antonio Gabriel Ravelo García, Arguente

Fábio Rúben Silva Mendonça, Vogal

Funchal, September 2026

Abstract

Sleep apnea affects nearly one billion people worldwide, yet diagnosis still relies on polysomnography, which is costly and complex. Wearable photoplethysmographic (PPG) sensors could support large-scale screening, but current methods mostly estimate respiratory rate or detect apnea within a single dataset, and phenotype analysis and cross-sensor generalization remain limited. This work presents a deep learning pipeline for sleep apnea detection, severity estimation and phenotype classification from PPG signals.

A CNN-LSTM model first showed that respiratory information can be extracted from PPG, with mean absolute errors of 1.70 and 1.80 breaths per minute on two held-out public test sets (BIDMC and CapnoBase). The CoTELMo architecture, a CNN-BiLSTM model trained on event-level annotations from 2,056 patients of the Multi-Ethnic Study of Atherosclerosis (MESA), then screened moderate-to-severe apnea with an Area Under the Receiver Operating Characteristic Curve (AUC) of 0.909, and a two-stage regression estimated the apnea-hypopnea index with a Pearson correlation of 0.853.

On the Cleveland Family Study (N=324), performance fell to an AUC of 0.568 without adaptation, while fine-tuning with a frozen backbone and patient-level labels recovered an AUC of 0.832. In an exploratory pilot with a commercial wearable ring (N=46, of which 12 were pathological), direct inference reached an AUC of 0.917 for severe obstructive apnea, and CoTELMo-derived features separated obstructive from central apnea in all 12 pathological cases. These pilot results require validation in larger cohorts.

The results suggest that PPG can support apnea screening and severity estimation, but that transfer to new sensors and populations requires supervised adaptation and local validation.

Keywords: Deep Learning · Photoplethysmography · Sleep Apnea · Transfer Learning · Phenotype Classification · Wearable Devices

Resumo

A apneia do sono afeta cerca de mil milhões de pessoas em todo o mundo, mas o diagnóstico continua a depender da polissonografia, um exame dispendioso e complexo. Os sensores fotopletismográficos (PPG) de dispositivos vestíveis podem apoiar o rastreio em larga escala, mas os métodos atuais focam-se sobretudo na frequência respiratória ou na deteção de apneia num único conjunto de dados, e a análise fenotípica e a generalização entre sensores continuam limitadas. Esta dissertação apresenta um pipeline de aprendizagem profunda para deteção de apneia do sono, estimativa de severidade e classificação fenotípica a partir de sinais PPG.

Um modelo CNN-LSTM mostrou primeiro que é possível extrair informação respiratória do PPG, com erros absolutos médios de 1,70 e 1,80 respirações por minuto em dois conjuntos de teste públicos reservados (BIDMC e CapnoBase). A arquitetura CoTELMo, um modelo CNN-BiLSTM treinado com anotações ao nível do evento de 2.056 pacientes do Multi-Ethnic Study of Atherosclerosis (MESA), rastreou apneia moderada a severa com AUC de 0,909, e uma regressão em duas fases estimou o índice de apneia-hipopneia com uma correlação de Pearson de 0,853.

No Cleveland Family Study (N=324), o desempenho desceu para AUC 0,568 sem adaptação, enquanto o ajuste fino com backbone congelado e rótulos ao nível do paciente recuperou AUC 0,832. Num piloto exploratório com um anel vestível comercial (N=46, dos quais 12 patológicos), a inferência direta obteve AUC 0,917 para apneia obstrutiva severa, e as características derivadas do CoTELMo separaram a apneia obstrutiva da central nos 12 casos patológicos. Estes resultados do piloto requerem validação em coortes maiores.

Os resultados sugerem que o PPG pode apoiar o rastreio e a estimativa de severidade da apneia, mas que a transferência para novos sensores e populações exige adaptação supervisionada e validação local.

Palavras-chave: Aprendizagem Profunda, Fotopletismografia, Apneia do Sono, Transferência de Aprendizagem, Classificação Fenotípica, Dispositivos Vestíveis

Acknowledgments

I would like to express my gratitude to Professor Fábio Mendonça, my supervisor, for the opportunity to develop this thesis and for his continuous support, insightful suggestions, and encouragement throughout the research process.

I would like to thank Dr. Jeff Durmer, my co-supervisor, and the Absolute Rest team for providing access to the clinical dataset that made the wearable validation possible, and for the opportunity to collaborate on a project with real-world impact in sleep technology.

I would also like to acknowledge the National Sleep Research Resource (NSRR) for providing access to the MESA, CFS, HeartBEAT, BestAIR, and HomePAP datasets, which were essential for the training and validation stages of this work.

This work was carried out while I held a research grant from ARDITI (Agência Regional para o Desenvolvimento da Investigação, Tecnologia e Inovação), reference ARDITI-SIH-AI-2025-003-BI, within the Smart Islands Hub (SIH) project, funded by the European Union and by national funds through IAPMEI and the Recovery and Resilience Plan (application no. 761, call AAC no. 03/C16-i03/2022).

Finally, I am grateful to my family for their patience and support throughout this journey.

Table of Contents

List of Figures	xvii
List of Tables.....	xix
List of Acronyms	xxi
1 Introduction	1
1.1 Problem Statement	3
1.2 Research Questions and Objectives	4
1.3 Thesis Outline	5
1.4 Key Remarks	6
2 Background.....	7
2.1 Photoplethysmography	7
2.2 PPG-Derived Respiration	8
2.3 Sleep-Disordered Breathing	10
2.4 Deep Learning Fundamentals	12
2.5 Evaluation Metrics	16
2.6 Key Remarks	18
3 State of the Art	21
3.1 Respiratory Rate Estimation from PPG	21

3.2	Sleep Apnea Detection Using Machine Learning	23
3.3	Transfer Learning in Biosignal Processing	27
3.4	Summary and Identified Gaps	28
3.5	Key Remarks	29
4	Materials and Methods	31
4.1	Datasets	31
4.2	Signal Preprocessing	36
4.3	Classical Machine Learning Baseline	41
4.4	Respiratory Rate Regression	43
4.5	Sleep Apnea Detection with CoTELMo	47
4.6	Continuous AHI Estimation	54
4.7	Transfer Learning and Domain Adaptation	55
4.8	Phenotype Classification	58
4.9	Software and Hardware	61
4.10	Key Remarks	62
5	Results and Discussion	63
5.1	Classical Machine Learning Results	63
5.2	Deep Learning Classification on MESA	64
5.3	Multiclass Severity Grading	67
5.4	Continuous AHI Estimation	70

5.5	Respiratory Rate Regression	73
5.6	External Validation and Domain Adaptation	75
5.7	Clinical Validation on Absolute Rest	79
5.8	Phenotype Classification.....	81
5.9	Comparison with State of the Art	87
5.10	Key Remarks	89
6	Conclusion.....	91
6.1	Summary and Answers to Research Questions.....	91
6.2	Contributions.....	93
6.3	Limitations	94
6.4	Future Work.....	96
6.5	Key Remarks	97
	References	99

List of Figures

1	Pipeline overview	5
2	PPG respiratory modulation mechanisms (RIIV, RIFV, RIAV)	8
3	Apnea mechanisms	11
4	Illustrative 1D convolutional block	13
5	LSTM cell gating mechanism	14
6	Skip connections in CoTELMo	15
7	Stratified 5-fold cross-validation	18
8	AHI severity distribution across datasets	35
9	Signal preprocessing pipelines	36
10	Signal quality index filtering	39
11	Weak vs. strong labeling strategies	41
12	Multi-scale CNN-LSTM architecture for respiratory rate estimation	45
13	Ensemble strategy for respiratory rate estimation	46
14	Top-K aggregation strategy	49
15	CoTELMo architecture	50
16	Ensemble inference with top-K aggregation	53
17	Two-stage AHI estimation pipeline	55
18	Frozen backbone transfer learning strategy	57

19	Deep learning model progression on MESA	66
20	Multiclass severity confusion matrix on MESA	69
21	Continuous AHI estimation on MESA: predicted vs. true.	71
22	Bland-Altman analysis of the AHI estimates on MESA	72
23	Auxiliary Ridge coefficient analysis	72
24	Domain adaptation strategy comparison on CFS	77
25	External validation and domain adaptation.	78
26	Silent phenotype: HV Load by severity	82
27	Phenotype classification confusion matrices	86

List of Tables

1	Deep learning methods for respiratory rate estimation from PPG.....	22
2	Summary of representative studies	28
3	MESA AHI severity distribution	33
4	CFS AHI severity distribution	33
5	Dataset summary	35
6	Classical ML signal comparison on MESA.....	63
7	Deep learning progression on MESA	65
8	Patient-level operating points at $AHI \geq 15$	67
9	Multiclass severity confusion matrix (table)	68
10	AHI regression per severity group	70
11	Loss function comparison for respiratory rate estimation	73
12	Respiratory rate model development milestones	74
13	Respiratory rate estimation comparison	74
14	Direct inference on external datasets	76
15	Domain adaptation results on CFS	76
16	ExtraTrees phenotype confusion matrix	83
17	Comparison with published methods	87

List of Acronyms

AASM	American Academy of Sleep Medicine
AC	Alternating Current
ADASYN	Adaptive Synthetic Sampling
AHI	Apnea-Hypopnea Index
ANOVA	Analysis of Variance
ASV	Adaptive Servo-Ventilation
AUC	Area Under the Receiver Operating Characteristic Curve
BCE	Binary Cross-Entropy
BIDMC	Beth Israel Deaconess Medical Center
BiLSTM	Bidirectional Long Short-Term Memory
BN	Batch Normalization
CFS	Cleveland Family Study
CNN	Convolutional Neural Network
CORAL	Correlation Alignment
CoTELMo	Convolutional TELMo
CPAP	Continuous Positive Airway Pressure
CPU	Central Processing Unit
CSA	Central Sleep Apnea
CUDA	Compute Unified Device Architecture
CV	Cross-Validation
DC	Direct Current
DL	Deep Learning
ECG	Electrocardiography
EDR	ECG-Derived Respiration
EEG	Electroencephalography
FC	Fully Connected
FFT	Fast Fourier Transform
GAP	Global Average Pooling
GPU	Graphics Processing Unit
HRV	Heart Rate Variability
ICC	Intraclass Correlation Coefficient
ICU	Intensive Care Unit
LED	Light-Emitting Diode

LOO	Leave-One-Out
LOOCV	Leave-One-Out Cross-Validation
LR	Logistic Regression
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MESA	Multi-Ethnic Study of Atherosclerosis
ML	Machine Learning
NSRR	National Sleep Research Resource
ODI	Oxygen Desaturation Index
OSA	Obstructive Sleep Apnea
PPG	Photoplethysmography
PPI	Peak-to-Peak Interval
PSG	Polysomnography
PVC	Premature Ventricular Contraction
RBF	Radial Basis Function
ReLU	Rectified Linear Unit
RIAV	Respiratory-Induced Amplitude Variation
RIFV	Respiratory-Induced Frequency Variation
RIIV	Respiratory-Induced Intensity Variation
RIPV	Respiratory-Induced Pulse-width Variation
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
RSA	Respiratory Sinus Arrhythmia
SDB	Sleep-Disordered Breathing
SHHS	Sleep Heart Health Study
SMOTE	Synthetic Minority Over-sampling Technique
SpO₂	Peripheral Capillary Oxygen Saturation
SQI	Signal Quality Index
SVR	Support Vector Regression
T90	Time with SpO ₂ below 90%
TELMo	Temporal ELMo
TPE	Tree-structured Parzen Estimator
VRAM	Video Random Access Memory

1 Introduction

Sleep is a fundamental biological process that occupies approximately one-third of human life. During sleep, the body performs critical restorative functions including memory consolidation, hormonal regulation, immune system maintenance, and cardiovascular repair [1]. Disruptions to normal sleep stages are linked to a wide range of harmful effects, from impaired cognitive function and reduced workplace productivity to increased risk of chronic diseases including diabetes, obesity, and neurodegenerative conditions [2]. Among the various conditions that compromise sleep quality, sleep apnea is particularly difficult to detect, as the repeated breathing cessations occur during sleep and are typically noticed only by bed partners, if at all.

Sleep apnea affects nearly one billion people worldwide, with Obstructive Sleep Apnea (OSA) being the most prevalent form [3]. This condition is characterized by repeated episodes of partial or complete obstruction of the upper airway during sleep, leading to intermittent hypoxia and sleep fragmentation. Each apnea event triggers a cascade of acute physiological responses (oxygen desaturation, sympathetic activation, intrathoracic pressure swings, and cortical arousal) that, when repeated dozens or hundreds of times per night over months and years, contribute to cumulative damage across multiple organ systems. The health consequences are well documented and include hypertension, cardiovascular disease, stroke, type 2 diabetes, cognitive impairment, depression, and increased all-cause mortality [4, 5]. Untreated severe OSA is associated with two to three times higher risk of motor vehicle accidents due to excessive daytime sleepiness [6], and the economic burden of undiagnosed sleep apnea in the United States alone has been estimated at \$149.6 billion per year in healthcare costs, lost productivity, and accident-related expenses [7].

Despite its prevalence and serious consequences, the vast majority of cases remain undiagnosed. Estimates suggest that 80–90% of moderate-to-severe OSA cases in the general population have never been identified [8]. This diagnostic gap is driven by the limitations of polysomnography (PSG), the standard diagnostic procedure. A single PSG study is expensive, labour-intensive, and time-consuming, requiring technician-attended overnight monitoring across more than 20 channels [9]. At this throughput, PSG cannot scale to meet the demand of population-level screening. The mismatch between the estimated one billion affected individuals and the global capacity for

PSG studies creates an urgent need for accessible, scalable screening tools that can identify individuals who would benefit from formal diagnostic evaluation.

The consequences of this diagnostic gap extend beyond individual patients. Undiagnosed OSA increases healthcare utilization through emergency visits for cardiovascular events, metabolic complications, and motor vehicle accidents that could have been prevented with appropriate treatment. From a public health perspective, the economic argument for scalable screening is clear: if a low-cost screening tool could identify even 5% of the estimated 900 million undiagnosed cases worldwide, the reduction in downstream healthcare costs would far exceed the cost of screening. The challenge is to develop such a tool using signals that are already available from consumer wearable devices, avoiding the need for additional clinical infrastructure.

Sleep monitoring devices are classified into four types based on their complexity. Type 1 corresponds to full in-laboratory PSG with a technician present. Type 2 systems are portable but still require a similar number of channels. Type 3 devices are limited-channel monitors that typically record airflow, respiratory effort, and oximetry. Type 4 devices are the simplest, recording only one or two channels, usually pulse oximetry or photoplethysmography (PPG), without electroencephalography, airflow, or respiratory effort measurements [10]. This last category is relevant because PPG sensors are already embedded in consumer wearable devices such as smartwatches, fitness bands, and pulse oximeter rings [11]. If the PPG signal alone could provide clinically useful information about sleep apnea, it would open the door to large-scale screening using hardware that millions of people already wear.

Beyond heart rate monitoring, the PPG signal encodes respiratory activity through subtle modulations in its intensity, amplitude, and frequency components [12]. These modulations arise from the mechanical coupling between respiration and the cardiovascular system, and they carry information about both respiratory rate and the presence of abnormal breathing patterns. The PPG waveform reflects autonomic nervous system activity indirectly through pulse rate variability and pulse morphology changes, which are altered during apnea events due to cyclic hypoxia, sympathetic activation, and arousal. However, extracting this information reliably is challenging due to the dominance of the cardiac component and susceptibility to motion artifacts [13]. Cardiac arrhythmias may also produce PPG morphology changes that motivate the dedicated PVC

robustness analysis performed in this work. Respiratory modulations in the PPG waveform are substantially smaller than the cardiac component, making their extraction non-trivial [14]. These challenges have limited the clinical adoption of PPG-based respiratory monitoring, despite the clear practical advantages of wearable sensors.

1.1 Problem Statement

The existing literature on PPG-based respiratory monitoring presents several limitations that this work aims to address. Many published methods focus on respiratory rate estimation, treating respiration as a scalar value in breaths per minute [15], or evaluate apnea detection on a single dataset. While respiratory rate is a useful vital sign, it does not capture the nature of the breathing abnormality. A patient with Central Sleep Apnea (CSA), where the brain fails to send signals to the respiratory muscles, and a patient with OSA, where the airway is physically blocked despite continued effort, may have similar respiratory rate values but require different treatments [16]. OSA is typically managed with Continuous Positive Airway Pressure (CPAP), while CSA may require Adaptive Servo-Ventilation (ASV), which provides dynamic breath-by-breath pressure support [17]. Prior work has distinguished central and obstructive events and screened central apnea burden from finger PPG [18]; however, three-class subject-level discrimination using single-wavelength wearable-ring PPG remains insufficiently studied.

The second limitation concerns cross-dataset generalization. Most studies report performance on a single dataset, often using internal cross-validation, without testing whether the model works on data from different sensors, populations, or acquisition settings [19]. When models trained on one dataset are applied to another, performance often degrades due to domain shift, which refers to systematic differences in signal characteristics between datasets caused by hardware, recording protocols, or patient demographics. Understanding and addressing this domain shift is necessary for any method that aims to work in real clinical settings where the specific sensor and patient population cannot be controlled. The problem is compounded by the diversity of PPG sensors in the consumer market: each manufacturer uses different light-emitting diode (LED) wavelengths, photodetector types, sampling rates, and analog filtering, producing signals that may look similar to the naked eye but differ in subtle ways that affect machine learning model performance.

The third limitation is the absence of transfer learning strategies specifically designed for PPG-based apnea detection across sensors and populations. Transfer learning has been applied in other biosignal domains, such as sleep staging from ECG [20,21], but its application to PPG-based apnea detection remains underexplored. Large-scale PSG datasets with event-level annotations exist (e.g., MESA with over 2,000 patients) and could serve as a foundation for training models that then adapt to smaller clinical datasets from wearable devices, but this source-to-target adaptation approach has not been investigated in the literature. The key challenge is determining how much of the source-domain knowledge transfers to a new sensor and what adaptation strategy is needed to recover the remaining performance gap.

1.2 Research Questions and Objectives

To address the gaps identified in the previous section, four research questions were formulated. RQ1 asks whether deep learning models can extract reliable respiratory rates from single-lead PPG signals in the presence of real-world noise and artifacts. RQ2 asks whether a deep learning architecture trained on large-scale polysomnographic data can achieve clinically useful sleep apnea screening from PPG alone. RQ3 asks whether the learned representation transfers across different sensors and patient populations, and what adaptation strategies are effective when it does not. RQ4 asks whether it is possible to distinguish OSA and CSA phenotypes using only PPG morphological features.

To answer these questions, six objectives were defined for this work. The first is to estimate respiratory rate from single-lead PPG signals with clinically acceptable accuracy using deep learning, with held-out evaluation across two clinical cohorts. The second is to develop a deep learning architecture for sleep apnea screening from PPG trained on a large-scale polysomnographic cohort with event-level annotations. The third is to evaluate cross-dataset generalization and establish adaptation strategies for deployment across different sensors and populations. The fourth is to validate the complete pipeline on a commercial wearable PPG device. The fifth is to distinguish apnea phenotypes (OSA and CSA) from PPG morphological features alone. The sixth is to assess the robustness of the classification pipeline against cardiac arrhythmias commonly associated with sleep apnea [22].

1.3 Thesis Outline

This work is organized into six chapters. Chapter 1 presents the motivation, problem statement, research questions, objectives, and document structure. Chapter 2 provides the background on PPG physiology, respiratory modulation, sleep apnea phenotypes, deep learning, and evaluation metrics. Chapter 3 reviews respiratory rate estimation, sleep apnea detection, and transfer learning in biosignal processing, and identifies the gaps addressed by this work. Chapter 4 describes the datasets, preprocessing, classical baselines, deep learning models, AHI estimation, transfer learning, and phenotype classification methods. Chapter 5 presents the results and comparisons with published methods. Chapter 6 summarizes the findings, contributions, limitations, and future directions. Figure 1 summarizes the complete pipeline.

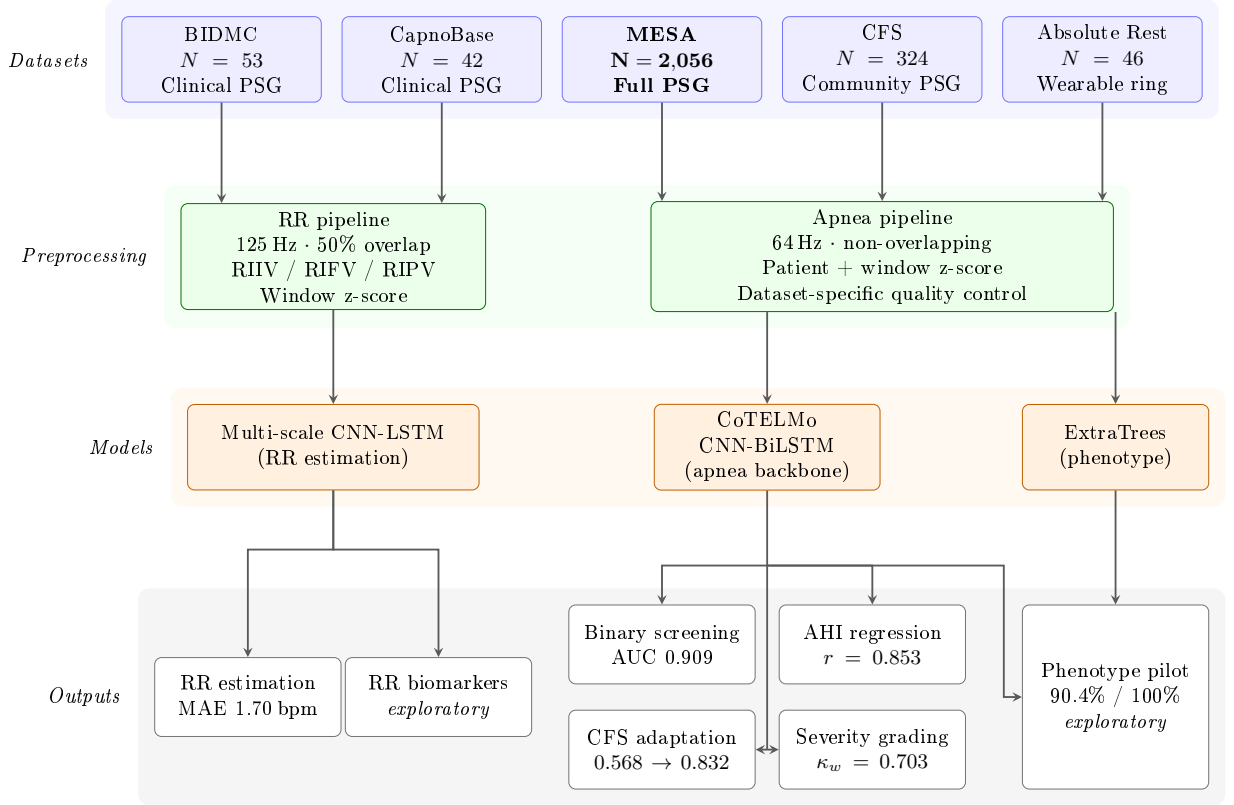


Fig. 1. Pipeline overview. Five datasets spanning different populations and sensors feed two preprocessing pipelines: respiratory rate at 125 Hz and apnea analysis at 64 Hz. CoTELMo supports binary screening, continuous AHI estimation, severity grading, CFS adaptation, and the exploratory phenotype pilot alongside ExtraTrees.

1.4 Key Remarks

Sleep apnea is a highly prevalent condition with serious health consequences, yet it remains largely underdiagnosed because the standard diagnostic method is expensive and inaccessible. Wearable PPG sensors offer a scalable alternative for screening, but phenotype evidence remains limited, cross-dataset generalization is poorly characterized, and practical transfer strategies across devices and populations are scarce. This work addresses these gaps through a pipeline that combines deep learning for apnea detection and severity estimation, transfer learning for cross-sensor adaptation, and morphological feature classification for an exploratory three-class phenotype analysis. The pipeline is evaluated across five datasets spanning different populations, sensors, and clinical settings, from large-scale epidemiological studies to a commercial wearable ring designed for elite athletes.

2 Background

The PPG waveform is shaped by both cardiovascular and respiratory physiology, making it possible to extract clinically relevant breathing information from a sensor designed primarily for heart rate monitoring. The sections below cover the signal physiology underlying this capability, the clinical framework for sleep apnea phenotypes, the deep learning architectures used to exploit it, and the evaluation metrics applied throughout the work.

2.1 Photoplethysmography

PPG is a non-invasive optical technique that measures volumetric changes in blood flow in the peripheral vasculature [23]. The measurement principle is straightforward: a light source, typically an LED, illuminates the skin, and a photodetector measures the amount of light that is either transmitted through or reflected from the tissue. As the heart contracts, the arterial blood volume in the illuminated region increases, which absorbs more light and reduces the signal reaching the detector. When the heart relaxes, the blood volume decreases and more light is detected. This pulsatile modulation produces the characteristic PPG waveform.

The raw PPG signal is conventionally decomposed into two components [24]. The direct current (DC) component represents the relatively constant absorption by non-pulsatile tissue, venous blood, and baseline arterial blood volume. It varies slowly with respiration, vasomotor activity, and thermoregulation, but is effectively static on a beat-to-beat timescale. The alternating current (AC) component rides on top of the DC level and reflects the pulsatile arterial blood volume changes synchronous with the cardiac cycle. It is this AC component that provides the familiar pulse waveform used for heart rate estimation, and its morphological features (such as pulse amplitude, rise time, and dicrotic notch position) encode information about vascular compliance, autonomic tone, and peripheral perfusion [23].

PPG sensors operate in two configurations depending on the anatomical site. In transmission mode, the LED and photodetector are placed on opposite sides of a translucent tissue such as a fingertip or earlobe, and the detector measures light that has passed through the tissue. In reflectance mode, both the LED and detector are placed on the same side of the skin, and the detector measures light that has been backscattered by the tissue. Reflectance mode is the standard

configuration in wrist-worn devices and smart rings because it does not require a thin tissue between source and detector [24]. The Absolute Rest ring used in this work, for instance, records reflectance-mode PPG from the finger using a green LED at a variable rate, most commonly 71.4 Hz.

Clinical-grade pulse oximeters use two wavelengths (red and infrared) to compute blood oxygen saturation (SpO_2) from the red-to-infrared absorbance ratio [23]. Consumer wearable PPG sensors generally use a single green wavelength (around 530 nm), which provides strong pulsatile contrast but cannot compute SpO_2 because differential absorption between oxygenated and deoxygenated hemoglobin requires at least two wavelengths. The Absolute Rest ring records single-wavelength PPG, making PPG-only analysis the only viable approach.

2.2 PPG-Derived Respiration

Although the PPG signal is primarily a cardiovascular measurement, respiration modulates it through several distinct physiological mechanisms [12,25,26] used by the respiratory rate regression models described in Chapter 4. Figure 2 illustrates the three most commonly cited mechanisms.

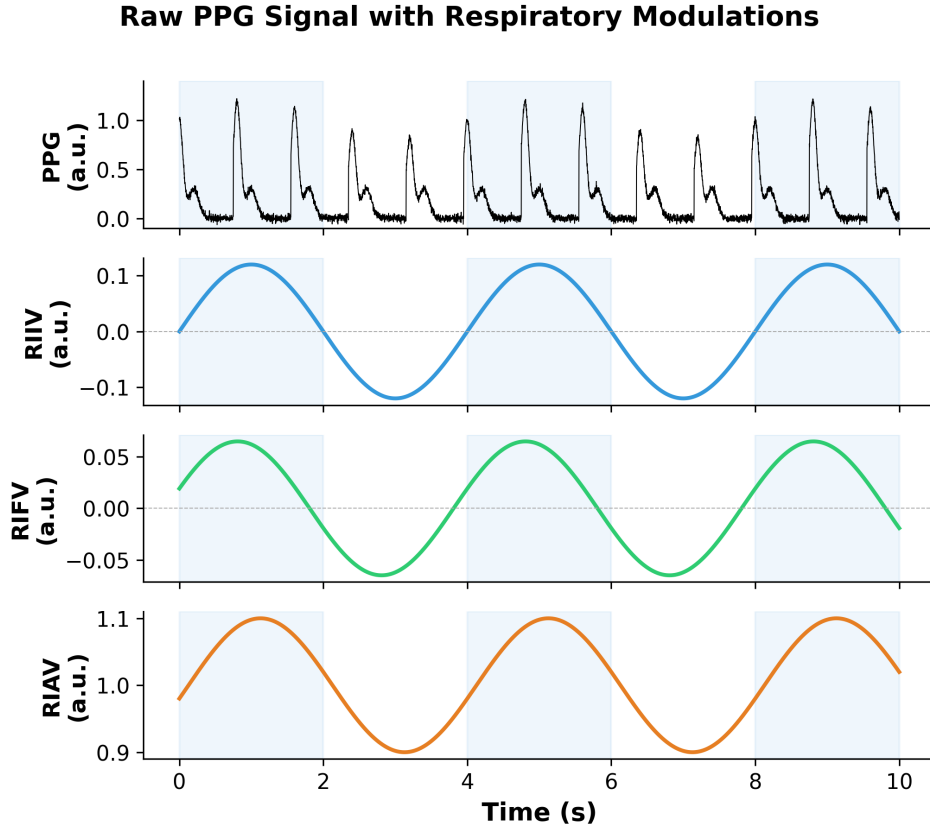


Fig. 2. Simulated PPG signal and respiratory-induced intensity (RIIV), frequency (RIFV), and amplitude (RIAV) variations. Light blue shading indicates respiratory cycle phases.

The first mechanism is Respiratory-Induced Intensity Variation (RIIV). During inspiration, intrathoracic pressure decreases, which increases venous return and transiently reduces peripheral blood volume. This produces a slow oscillation in the DC component of the PPG signal at the respiratory frequency, typically between 0.1 and 0.5 Hz for normal adult breathing rates of 6–30 breaths per minute [27]. RIIV is the strongest respiratory modulation in finger-based PPG. In the preprocessing pipeline described in Section 4.2, it is recovered from the slow baseline wander of the pulse train through interpolation of the detected peak positions, rather than by preserving the sub-0.5 Hz band of the raw signal directly.

The second mechanism is Respiratory-Induced Frequency Variation (RIFV), which arises from respiratory sinus arrhythmia (RSA), the physiological phenomenon whereby heart rate increases during inspiration and decreases during expiration due to vagal modulation of the sinoatrial node [25]. RIFV manifests as a periodic modulation of the peak-to-peak interval (PPI) in the PPG waveform. Unlike RIIV, which modulates the signal amplitude, RIFV modulates its timing. This mechanism is exploited by heart rate variability (HRV) features used in the classical machine learning baseline described in Section 4.3.

The third mechanism is Respiratory-Induced Amplitude Variation (RIAV). Changes in stroke volume caused by respiratory fluctuations in intrathoracic pressure produce cyclic variations in the amplitude of the PPG pulse wave [12]. During inspiration, decreased intrathoracic pressure increases left ventricular preload, leading to a transient increase in stroke volume and pulse amplitude. RIAV provides an independent respiratory signal that can be extracted by tracking the envelope of the AC component.

A closely related but distinct modulation is Respiratory-Induced Pulse-width Variation (RIPV). The same intrathoracic pressure changes that modulate pulse amplitude also affect left ventricular ejection time, producing cyclic variations in the temporal width of each cardiac pulse [14]. During inspiration, the increased venous return prolongs the ejection phase, widening the pulse; during expiration, the opposite occurs. RIPV is computed from the temporal distance between the systolic onset and the diastolic notch (or end of the pulse) for each cardiac cycle. In the respiratory rate regression pipeline described in Chapter 4, RIPV replaces RIAV as the third input channel because it proved more robust to baseline wander artifacts in the BIDMC and CapnoBase datasets.

These four mechanisms operate simultaneously and can be separated through appropriate signal processing. In the context of sleep apnea, their disruption during respiratory events, when airflow ceases or is reduced, produces characteristic patterns in the PPG signal that deep learning models can learn to recognize even without explicit feature extraction.

2.3 Sleep-Disordered Breathing

Sleep-Disordered Breathing (SDB) encompasses a spectrum of conditions characterized by abnormal respiratory patterns during sleep. The most clinically relevant forms are OSA, CSA, and hypopnea, which differ in their underlying mechanisms and therapeutic implications [28].

OSA is the most common form of SDB. An obstructive apnea event occurs when the upper airway collapses during sleep, completely blocking airflow for at least 10 seconds despite continued respiratory effort from the diaphragm and intercostal muscles [28]. The obstruction triggers a cascade of physiological responses: oxygen desaturation as gas exchange ceases, progressive bradycardia mediated by the diving reflex, increasing respiratory effort against the closed airway, and ultimately a cortical arousal that restores muscle tone and reopens the airway. This cycle repeats throughout the night, sometimes hundreds of times, and produces characteristic PPG signatures including amplitude drops (reflecting desaturation), heart rate variability changes, and transient tachycardia at the point of arousal.

CSA differs from OSA in that the cessation of airflow results from a failure of the central respiratory drive rather than airway obstruction [16]. During a central apnea event, neither airflow nor respiratory effort is present. CSA often occurs in patients with heart failure, stroke, or brainstem lesions, and may also emerge during treatment with opioid medications. The absence of respiratory effort reduces the intrathoracic pressure fluctuations that modulate PPG, so central and obstructive events may differ in effort-related morphology, although this pattern is not expected to be universal across patients.

Hypopnea events represent a partial reduction in airflow (at least 30% reduction lasting at least 10 seconds, accompanied by either oxygen desaturation of at least 3% or an arousal) rather than complete cessation [28]. Hypopneas are the most frequent respiratory events in most patients with

SDB and produce attenuated versions of the same PPG signatures as full apnea events. Figure 3 illustrates the key distinction between normal breathing, obstructive apnea, and central apnea.

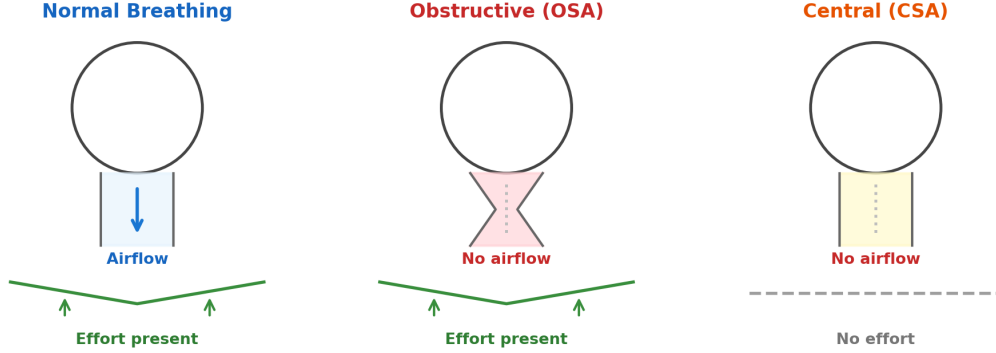


Fig. 3. Comparison of respiratory mechanisms in normal breathing, OSA, and CSA. In OSA, the airway collapses despite continued respiratory effort, producing intrathoracic pressure swings that may modulate PPG. In CSA, the airway remains open but neither effort nor airflow is present, reducing this effort-related modulation.

The clinical severity of SDB is quantified by the AHI, defined as the number of apnea and hypopnea events per hour of sleep. The American Academy of Sleep Medicine (AASM) defines four severity categories [29]: Normal ($AHI < 5$), Mild ($5 \leq AHI < 15$), Moderate ($15 \leq AHI < 30$), and Severe ($AHI \geq 30$). These thresholds have direct therapeutic implications: patients with Mild SDB may be managed with positional therapy or oral appliances, while Moderate and Severe cases typically require CPAP therapy for OSA or ASV for CSA [17,30]. The distinction between OSA and CSA is clinically critical because CPAP, while effective for obstructive events, can worsen central events in some patients by suppressing the hypercapnic ventilatory drive. In clinical practice, phenotype determination currently requires full PSG with respiratory effort bands (thoracic and abdominal) to distinguish between obstructive events (where effort is present but airflow is absent) and central events (where both effort and airflow are absent). This distinction cannot be made from pulse oximetry, ECG, or airflow signals alone, which is why PSG remains the gold standard despite its cost and logistical limitations.

The AHI metric, while clinically established, has known limitations. It treats all events as equivalent regardless of their duration, depth of desaturation, or degree of arousal. A patient with 20 events per hour lasting 11 seconds each and producing 3% desaturations receives the same AHI as a patient with 20 events per hour lasting 45 seconds each and producing 12% desaturations,

despite the latter experiencing far greater physiological stress. Alternative severity indices such as the Oxygen Desaturation Index (ODI), time with SpO_2 below 90% (T90), and event-duration-weighted metrics have been proposed but have not replaced AHI in clinical practice. For the purposes of this work, AHI is used as the reference severity metric for the sleep-apnea datasets because it provides direct comparability with prior literature.

2.4 Deep Learning Fundamentals

This section introduces the deep learning building blocks used in the architectures developed in this work: convolutional neural networks for feature extraction from raw signals, recurrent networks for temporal modeling, and regularization techniques for generalization.

Convolutional Neural Networks (CNNs) learn hierarchical feature representations through layers of learnable filters that are convolved across the input [31, 32]. While CNNs were originally developed for two-dimensional image processing, their one-dimensional variant is well suited for time-series analysis, where the input is a sequence of scalar values sampled at regular time intervals. In one-dimensional form, a convolutional layer applies a set of K filters of length L to an input sequence, producing K feature maps. Each filter computes a local weighted sum at every position along the sequence, and a non-linear activation function (typically ReLU, $f(x) = \max(0, x)$) is applied element-wise. By stacking multiple convolutional layers with increasing filter counts (e.g., $32 \rightarrow 64 \rightarrow 128 \rightarrow 256$), the network learns features at progressively larger temporal scales: early layers capture local morphological patterns such as pulse shape and peak characteristics, while deeper layers capture multi-beat patterns spanning several seconds. Pooling layers reduce the spatial dimension of the feature maps by replacing local regions with summary statistics, typically the maximum value (max pooling). Batch normalization (BN) layers [33] normalize the activations within each mini-batch to stabilize and accelerate training. When the full sequence of feature maps is to be condensed into a fixed-length vector, Global Average Pooling (GAP) computes the mean across the temporal dimension for each filter, producing a vector of length K that represents the entire input window. In this work, 1D CNNs serve as feature extractors for both the standalone CNN baseline and the convolutional front-end of the CoTELMo architecture. The convolutional layers transform raw PPG windows of 1,920 samples (30 seconds at 64 Hz) into compact feature

sequences that can be processed by subsequent recurrent layers or classification heads. Figure 4 illustrates the structure of a single convolutional block.

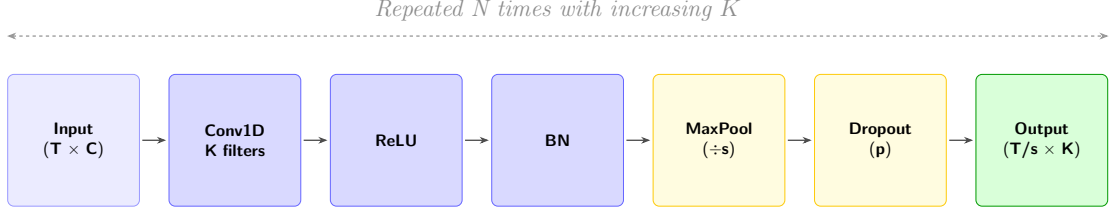


Fig. 4. Illustrative structure of a 1D convolutional block. Learnable filters of length L produce K feature maps, followed by activation, normalization, pooling, and optional dropout. The exact ordering and regularization differ between the respiratory-rate and CoTELMo architectures.

Recurrent Neural Networks (RNNs) process sequential data by maintaining a hidden state vector that is updated at each timestep based on the current input and the previous hidden state. However, vanilla RNNs suffer from the vanishing gradient problem, where gradients decay exponentially during backpropagation through time, preventing the network from learning long-range dependencies [34]. The Long Short-Term Memory (LSTM) architecture addresses this limitation through a gating mechanism that controls the flow of information into, within, and out of a memory cell [34]. At each timestep t , the LSTM computes three gates: the forget gate f_t determines which information to discard from the cell state, the input gate i_t determines which new information to store, and the output gate o_t determines which parts of the cell state to expose as the hidden state. The cell state c_t acts as a conveyor belt that carries information across many timesteps with minimal gradient degradation, allowing the network to capture dependencies over sequences of hundreds of steps. The Bidirectional LSTM (BiLSTM) extends the LSTM by processing the input sequence in both the forward and backward directions simultaneously [35]. The forward LSTM captures dependencies from past to future, while the backward LSTM captures dependencies from future to past. At each timestep, the outputs of both directions are concatenated, giving the network access to the full temporal context. In the CoTELMo architecture, BiLSTM layers process the compressed feature sequence from the CNN, capturing temporal patterns that span the entire 30-second window, such as the characteristic progression of normal breathing followed by cessation, increasing effort, desaturation, and arousal that defines an apnea event. Figure 5 illustrates the internal gating mechanism of an LSTM cell.

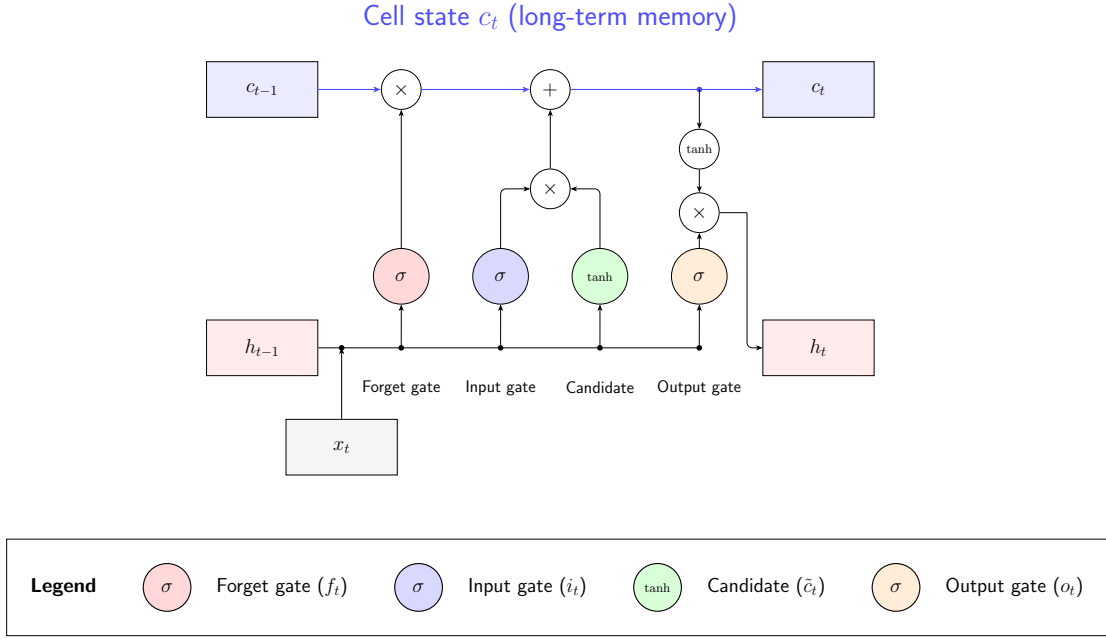


Fig. 5. Internal architecture of an LSTM cell. The forget gate (f_t) controls which information to discard from the cell state, the input gate (i_t) selects new information to store, and the output gate (o_t) determines what to expose as the hidden state h_t . The cell state c_t acts as a long-term memory conveyor belt with minimal gradient degradation.

Skip connections, also known as residual connections [36], provide direct pathways for information to bypass one or more layers in the network. In the context of the TELMo architecture from which CoTELMo derives [37], skip connections allow the classification head to receive features at multiple temporal resolutions: the raw CNN features (fine-grained, local), the first BiLSTM output (medium-range context), and the second BiLSTM output (full-window context). These multi-resolution features are concatenated and jointly processed by the classifier, enabling it to combine local pulse morphology with broader temporal patterns. Figure 6 illustrates the skip connection mechanism as used in the CoTELMo architecture.

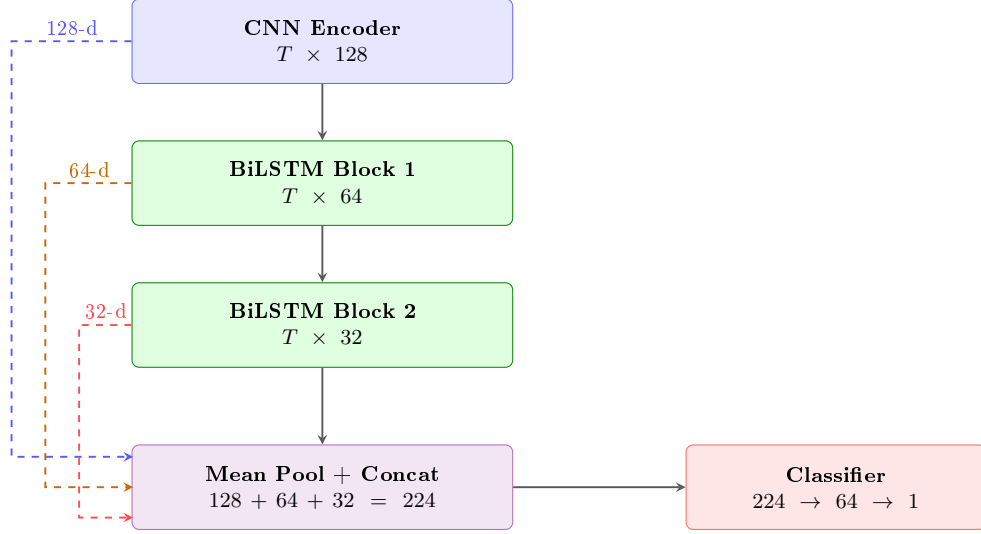


Fig. 6. Multi-resolution feature fusion via skip connections in the CoTELMo architecture. Features from the CNN encoder (128-dim), BiLSTM Block 1 (64-dim), and BiLSTM Block 2 (32-dim) are concatenated after mean pooling into a 224-dimensional vector, giving the classifier access to features at multiple temporal scales.

Overfitting is a persistent challenge in medical signal processing, where datasets are often small relative to model capacity. Dropout [38] is a regularization technique that randomly sets a fraction p of activations to zero during training, forcing the network to learn redundant representations that are robust to the loss of individual neurons. Weight decay (L2 regularization) adds a penalty proportional to the squared magnitude of the weights to the loss function, discouraging large weight values that overfit to training data. Both techniques are used throughout the architectures in this work. Transfer learning is the practice of reusing knowledge learned on one task or domain to improve performance on a different but related task [39]. In this work, transfer learning is employed in two contexts. First, the CoTELMo model trained on MESA polysomnographic data is adapted to the CFS and Absolute Rest datasets by freezing the learned feature extraction layers and retraining only the classification head on the target domain. This approach assumes that the convolutional and recurrent layers have learned generalizable representations of PPG patterns, while the classification head needs recalibration for the specific signal characteristics of each sensor.

The choice between freezing the backbone and full fine-tuning involves a bias-variance trade-off that depends on the size of the target dataset. Full fine-tuning updates all parameters using target-domain data, which gives the model maximum flexibility to adapt but also maximum risk of overfitting when the target dataset is small. Freezing the backbone eliminates this risk by fixing the learned representations and only training a small number of new parameters (the classification

head), but it assumes that the source representations are sufficiently general to capture the relevant patterns in the target domain. When the target dataset contains fewer than a few hundred patients, as is the case for CFS (324 patients) and Absolute Rest (46 patients), freezing is generally preferred because the data is insufficient to reliably update the millions of backbone parameters.

The choice of which layers to freeze and which to retrain is informed by the nature of the domain shift: when the shift primarily affects signal amplitude and baseline characteristics (correctable by normalizing the decision boundary), freezing the entire backbone and training only a new head is sufficient. When the shift affects deeper signal properties (waveform morphology, frequency content), partial fine-tuning of the later convolutional layers may be necessary. In this work, the complete freeze strategy proved sufficient, suggesting that the MESA-CFS and MESA-Absolute Rest domain shifts are primarily decision-boundary shifts rather than representation-level shifts.

2.5 Evaluation Metrics

This section defines the evaluation metrics used to assess the performance of the models developed in this work. Different tasks require different metrics, and their proper interpretation is necessary for clinical relevance.

The AUC-ROC is the primary metric for binary classification tasks in this work [40]. The ROC curve plots sensitivity (true positive rate) against $1 - \text{specificity}$ (false positive rate) across all possible decision thresholds. The AUC summarizes this trade-off as a single scalar that usually ranges between 0.5 (random classifier) and 1.0 (perfect classifier), although values below 0.5 can happen if the classifier ranking is systematically worse than random. AUC is preferred over accuracy in medical screening because it is invariant to the choice of decision threshold and robust to class imbalance. Sensitivity (recall) measures the proportion of positive cases correctly identified, and specificity measures the proportion of negative cases correctly identified. In sleep apnea screening, high sensitivity is generally prioritized because a missed diagnosis (false negative) has greater clinical consequences than a false alarm (false positive), which would simply lead to a confirmatory PSG. Precision (positive predictive value) measures the proportion of predicted positives that are truly positive. The F1-score is the harmonic mean of precision and recall and provides a single metric that balances both.

For continuous AHI estimation, Pearson’s correlation coefficient (r) measures the linear association between predicted and true AHI values, ranging from -1 to $+1$. Spearman’s rank correlation (ρ) is a non-parametric alternative that measures the monotonic association. Mean Absolute Error (MAE) quantifies the average magnitude of prediction errors in events per hour, providing a directly interpretable clinical metric. Root Mean Square Error (RMSE) penalizes large errors more heavily than MAE. Bland-Altman analysis [41] is used to assess the agreement between two measurement methods, in this case the model’s AHI estimate and the PSG reference. The Bland-Altman plot displays the difference between the two methods against their mean, providing visual information about systematic bias (mean difference), random error (limits of agreement), and whether the error magnitude depends on the AHI value (proportional bias).

Cohen’s quadratic weighted kappa (κ_w) is used to evaluate multiclass severity grading (Normal, Mild, Moderate-to-Severe) [42]. Unlike simple accuracy, quadratic kappa penalizes misclassifications proportionally to their distance from the correct class: classifying a Severe patient as Normal is penalized far more heavily than classifying them as Moderate. The Landis and Koch scale [43] provides a qualitative interpretation: $\kappa_w < 0.20$ indicates slight agreement, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.81–1.00 almost perfect agreement.

Stratified k -fold cross-validation partitions the dataset into k folds such that each fold preserves the class distribution of the full dataset (Figure 7). All splits in this work are performed at the patient level, meaning that all windows from a given patient appear exclusively in either the training or test set, never both. This prevents data leakage from temporally correlated windows within the same recording and ensures that the reported metrics reflect true generalization to unseen patients.

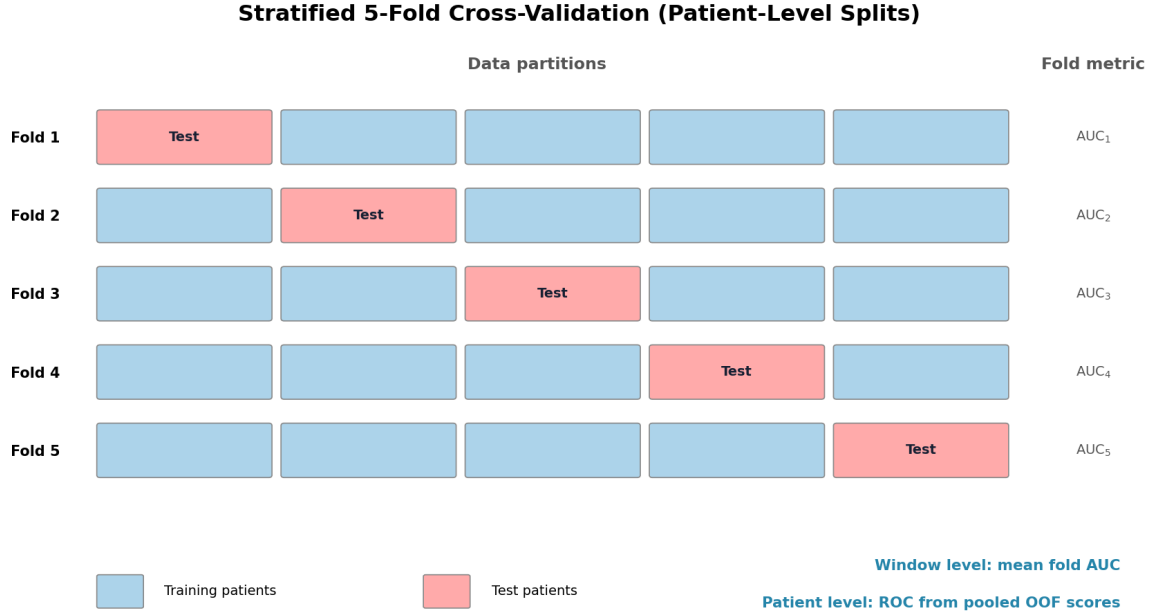


Fig. 7. Stratified 5-fold cross-validation with patient-level splits. All windows from each patient remain in one partition. Window-level AUC is summarized across folds; patient-level MESA AUC is computed once from pooled out-of-fold patient scores, so every patient is evaluated by a model that did not train on that patient.

Leave-One-Out Cross-Validation (LOOCV) is used when sample sizes are very small (e.g., the Absolute Rest phenotype classification with $N = 46$). In LOOCV, each patient serves as the test set exactly once, with all remaining patients forming the training set. This maximizes the use of available data for training but can produce high-variance estimates because each test fold contains only a single observation. The resulting performance estimate is nearly unbiased (the training set size $N - 1$ is close to the full dataset size N), but the variance across folds can be high, particularly when individual patients are outliers. For this reason, LOOCV results are reported with mean and standard deviation across random seeds when randomized components are involved (e.g., random forest seed, feature selection), and the consistency of results across all N iterations is examined as supporting evidence for reliability.

2.6 Key Remarks

The PPG signal, although primarily a cardiovascular measurement, encodes respiratory information through four distinct modulation mechanisms (RIIV, RIFV, RIAV, and RIPV) that operate at frequencies between 0.1 and 0.5 Hz. Sleep-disordered breathing disrupts these modulations in phenotype-specific ways: obstructive events produce characteristic effort-related oscillations and

desaturation signatures, while central events produce a quieter signal morphology due to the absence of respiratory effort. The deep learning architectures used in this work (1D CNNs for local feature extraction, BiLSTMs for temporal context, and skip connections for multi-resolution feature fusion) are designed to capture these patterns at multiple timescales directly from the raw signal. Transfer learning with frozen backbones and retrained classification heads provides a practical strategy for adapting models across sensors without requiring event-level annotations on the target device. The evaluation framework combines threshold-independent metrics (AUC), clinically interpretable error metrics (MAE in events per hour), and ordinal agreement measures (quadratic kappa) to provide an assessment of model performance across all clinical tasks.

3 State of the Art

Sleep monitoring has evolved from laboratory-based PSG toward portable and wearable solutions [44]. This chapter reviews the existing literature relevant to the three main tasks addressed in this work: respiratory rate estimation from PPG, sleep apnea detection using machine learning, and transfer learning in biosignal processing. The review is organized by task, with each section summarizing the main approaches, their results, and their limitations. The chapter concludes with a summary of the identified gaps that this work aims to address.

3.1 Respiratory Rate Estimation from PPG

The estimation of respiratory rate from PPG has been studied using both classical signal processing and deep learning approaches. As described in Chapter 2, respiration modulates the PPG waveform through four mechanisms (RIIV, RIFV, RIAV, and RIPV), and the challenge is to recover these subtle modulations from a signal dominated by the cardiac component. Classical methods and deep learning methods offer different trade-offs in terms of flexibility, generalization, and hardware requirements.

Classical methods typically extract one or more respiratory surrogates from the PPG waveform and then estimate the dominant respiratory frequency using spectral analysis or peak counting [14, 45]. The general pipeline involves beat detection, extraction of RIIV, RIFV, or RIPV by tracking pulse-to-pulse variations, bandpass filtering to the expected respiratory range (0.1–0.5 Hz), and frequency estimation via FFT, autoregressive modeling, or peak counting. Pimentel et al. [46] proposed a quality-weighted fusion framework that combines multiple respiratory estimates, achieving a MAE of 1.9 bpm on CapnoBase. Orphanidou et al. [47] developed signal quality indices for PPG that enabled automatic rejection of unreliable segments before respiratory analysis. Lazazzera and Carrault [48] compared classical breathing rate estimation methods on the CapnoBase database, confirming that fusion outperforms individual surrogates. The RRest toolbox by Charlton et al. [49] compared 314 combinations of extraction techniques, fusion strategies, and frequency estimation methods, establishing that no single method dominated across all datasets and that performance depended on signal quality, patient population, and recording conditions. A limitation shared by all classical methods is their dependence on accurate beat detection; when the cardiac cycle is irregular due to arrhythmias or motion artifacts, the extracted surrogates con-

tain spurious variations that corrupt the respiratory rate estimate. Furthermore, classical methods require manual selection of filter parameters, frequency bands, and fusion weights, all of which must be tuned for each specific dataset or recording condition. When applied to full-night recordings rather than the short (8-minute) controlled segments on which they were developed, classical methods face additional challenges: respiratory rate varies naturally during sleep (typically 12–20 breaths per minute in adults, with lower rates during deep sleep and higher rates during REM), and the signal quality fluctuates as the patient changes position, producing intermittent periods of poor optical coupling. In the studies reviewed here, classical methods have not been evaluated on full-night sleep recordings or shown to be robust to these real-world conditions.

Deep learning methods avoid the need for explicit beat detection and manual parameter selection by learning directly from the raw or minimally processed signal. Table 1 summarizes the main deep learning approaches for respiratory rate estimation from PPG.

Table 1. Deep learning methods for respiratory rate estimation using PPG-containing inputs. Values are MAE in breaths per minute (bpm); “—” indicates that the metric was not reported in a directly comparable form by the original study.

Method	Input	Architecture	BIDMC / CapnoBase
Bian 2020 [50]	PPG	ResNet	2.5 ^a
RespNet [51]	PPG	CNN + recon.	— ^b
RRWaveNet [52]	PPG	Multi-scale CNN	1.66 / 1.59
Chowdhury 2022 [53]	PPG	Lightweight CNN	0.77 / —
PPG2RespNet [54]	PPG	UNet++ skip	0.69 ^c / 0.11 ^c
Hwang 2023 [55]	PPG	ResNet/U-Net	1.26 ^d / —
Kumar 2022 [15]	PPG+ECG	Bi-LSTM-attn.	— ^e
Zhao 2023 [19]	PPG+ECG	Transformer	1.33 / 0.96

^a Combined MAE across BIDMC and CapnoBase; per-dataset values not reported.

^b RespNet reports respiratory-*signal* reconstruction error (MSE/correlation), not respiratory-*rate* MAE.

^c Reported as RR MAE; unusually low, with direct comparison limited by likely differences in the evaluation protocol.

^d Up to 3.1 bpm on the authors’ surgical-ICU dataset.

^e Kumar 2022 reports a best MAE of 0.24 bpm using PPG+ECG; per-dataset values for BIDMC and CapnoBase were not reported.

Bian et al. [50] proposed a ResNet that takes raw PPG segments as input and directly outputs respiratory rate. Ravichandran et al. [51] introduced RespNet, which reconstructs the full respiratory waveform from PPG before estimating rate. Osathitporn et al. [52] proposed RRRWaveNet, a multi-scale residual CNN that achieved a MAE of 1.66 bpm on BIDMC and 1.59 on Cap-

noBase. Shuzan et al. [54] introduced PPG2RespNet, which uses layered skip connections inspired by UNet++ to reconstruct respiratory signals across three datasets. An important distinction in the literature is between PPG-only and PPG+ECG methods. Kumar et al. [15] and Zhao et al. [19] reported that adding an ECG-derived respiratory channel alongside PPG reduced the error; the Transformer-based model of Zhao et al. reported the lowest errors (1.33 on BIDMC, 0.96 on CapnoBase) using PPG+ECG. However, the ECG requirement limits deployment on wrist or ring-based devices. This trade-off between performance and hardware requirements is directly relevant to the present work, which prioritizes PPG-only operation. Davies and Mandic [56] proposed a deep corr-encoder for respiratory waveform extraction, while Hwang et al. [55] evaluated multiple architectures in a surgical ICU setting. Chowdhury et al. [53] developed a lightweight model for resource-constrained wearable deployment. Liu et al. [57] noted in a recent review that deep learning approaches were beginning to match classical methods, though their dependence on large training datasets remained a limitation.

Signal quality assessment remains an important preprocessing step, as PPG signals are easily corrupted by motion artifacts and body movements [58]. A common limitation across all reviewed studies is the reliance on short clinical recordings (typically 8 minutes from BIDMC or CapnoBase) from controlled hospital environments. Whether the models generalize to full-night recordings, ambulatory settings, or wearable sensors with different signal characteristics has not been evaluated. Furthermore, none of these studies investigated whether the extracted respiratory information could serve as a biomarker for sleep-disordered breathing, which is the connection explored in this work.

3.2 Sleep Apnea Detection Using Machine Learning

Sleep apnea detection using machine learning has been approached from multiple signal modalities, including ECG, SpO₂, and PPG. The choice of signal determines the available information: ECG captures cardiac rhythm disturbances, SpO₂ directly measures the oxygen desaturation that follows airway obstruction, and PPG captures a combination of cardiac, respiratory, and autonomic modulations through a single optical sensor. This section reviews each modality, with emphasis on PPG-based methods.

ECG-based methods have the longest history [59]. Urtnasan et al. [60] applied a CNN to single-lead ECG for binary detection of obstructive apnea events, reporting an F1-score of 0.96

on their own single-cohort dataset ($N=82$). However, this result must be interpreted cautiously given the small, single-centre dataset. More broadly, the Physionet Apnea-ECG dataset, which has served as the primary benchmark for ECG-based methods since its release in 2000, uses a binary labeling scheme (apnea vs. normal per minute) that does not distinguish between event types and does not reflect the severity grading used in clinical practice. Its small size (35 training and 35 test recordings) further means that accuracies reported on it may not generalize to broader populations. Bahrami and Forouzanfar [61] compared classical and deep learning approaches for ECG-based detection and found that CNNs and LSTMs consistently outperformed traditional feature-based methods. Bozkurt et al. [62] combined ECG features with hybrid machine learning for respiratory event detection. Mostafa et al. [63] conducted a review of deep learning approaches and found that most studies used ECG or SpO_2 as input, with CNN being the most common architecture, and identified a lack of standardization in evaluation protocols.

SpO_2 -based methods exploit the direct relationship between oxygen desaturation and apnea events. When airflow ceases, blood oxygen saturation drops within 10–30 seconds, producing a characteristic desaturation-reoxygenation pattern. Deviaene et al. [64] developed automatic SpO_2 -based screening using statistical features. Gutiérrez-Tobal et al. [65] applied machine learning to oximetry-derived features to estimate sleep apnea severity. Leino et al. [66] applied a neural network to nocturnal SpO_2 signals for sleep apnea severity screening in stroke and transient ischaemic attack patients, without manual signal annotation. Levy et al. [67] introduced OxiNet, a deep learning model applied to single-channel oximetry data from 12,923 polysomnography recordings across six databases, reporting an intraclass correlation of 0.92–0.96 for AHI estimation and a macro F1-score of 0.75–0.84 for severity classification. OxiNet represents the closest methodological comparison to the present work in terms of scale and multi-site evaluation, though it operates on SpO_2 rather than raw PPG. The main limitation of SpO_2 -based methods is the hardware requirement: SpO_2 requires dual-wavelength optical sensors (red and infrared LEDs), which are not available in devices that use only green LEDs for heart rate monitoring. Additionally, desaturation may be absent, delayed, or attenuated in mild cases and in CSA where respiratory effort is absent.

PPG-based methods are less mature but of increasing interest due to the ubiquity of single-wavelength PPG sensors in consumer wearables. Hayano et al. [68] used heart rate variability

features from a PPG wrist watch to detect moderate-to-severe OSA in 41 patients, reporting an AUC of 0.84. Jiang et al. [69] applied a CNN to raw PPG signals from 59 patients and reported an accuracy exceeding 90%. At a larger scale, Jothi et al. [70] trained deep networks for OSA event detection using fingertip PPG from 1,375 subjects. Fonseca et al. [71] showed that cardiorespiratory features carry sleep-related information, motivating the search for equivalent markers in PPG. Papini et al. [72] used 212 features derived from wrist-worn PPG in a deep learning model to estimate AHI, achieving a correlation of $r = 0.61$ with the PSG reference. Wei et al. [73] proposed MS-net, a multiscale convolutional network for apnea detection from PPG, achieving 82.0% per-segment and 93.6% per-recording accuracy. Massie et al. [18] demonstrated central-versus-obstructive respiratory-event classification and patient-level central-apnea screening using finger PPG, establishing an important phenotype-related precedent. Korkalainen et al. [74] demonstrated automatic sleep staging from finger PPG in 894 patients, showing that PPG also supports multi-stage sleep classification at scale. Mendonça et al. [75] developed a wireless oximetry-based device for portable sleep apnea detection. Casal et al. [76] classified sleep-wake stages from pulse oximetry signals using recurrent neural networks. Olsen et al. [77] applied deep transfer learning to raw wrist accelerometer and PPG data from consumer sleep technologies for combined sleep staging and apnea detection.

A pattern that emerges from the PPG-based literature is the tension between the information content of the signal and the practical constraints of the hardware. The PPG waveform encodes respiratory, cardiac, and autonomic information through the modulation mechanisms described in Chapter 2, but these modulations are subtle and can be corrupted by the very conditions that make PPG attractive for screening: uncontrolled recording environments, variable sensor placement, and consumer-grade hardware with lower signal-to-noise ratios than clinical PSG sensors. The studies that report the highest accuracy tend to use clinical-grade pulse oximeters in controlled settings, while those using consumer wearables report lower and more variable performance. This gap between laboratory and real-world performance is not unique to PPG-based apnea detection, but it is particularly relevant because the clinical value of PPG-based screening depends entirely on its ability to work outside the laboratory.

These studies collectively demonstrate that PPG signals contain clinically relevant information about sleep-disordered breathing, but the field remains fragmented, with no unified pipeline that addresses detection, severity estimation, and phenotype classification in a single framework.

The growing adoption of PPG-enabled smartwatches has been driven by early successes in other cardiac monitoring tasks, and wearable sensors are increasingly recognized as a route to large-scale sleep monitoring that PSG cannot provide [78]. Tison et al. [79] demonstrated passive atrial fibrillation detection using a commercially available smartwatch, establishing the clinical potential of wrist-worn PPG sensors. In the sleep apnea domain, the Belun Ring, an FDA-cleared PPG-based ring, uses deep learning for OSA detection, reporting a sensitivity of 0.91 and specificity of 0.88 for moderate-to-severe OSA ($AHI \geq 15$) [80]. Lazazzera et al. [81] developed a wearable device combining PPG with blood pressure estimation. Wang et al. [82] proposed a dual-modal wearable integrating a piezoelectric nanogenerator with PPG. These devices demonstrate clinical feasibility, although phenotype evidence remains limited and dependent on specific finger-based acquisition systems.

Two recent reviews provide a picture of the current state. Abd-Alrazaq et al. [83] examined 38 studies using wearable AI for sleep apnea detection and reported a pooled accuracy of 89.3% for event-level detection, noting that sensitivity for apnea events (79.3%) lagged behind specificity for normal events (94.7%). Osa-Sanchez et al. [84] covered 28 studies from 2020 to 2024 and found that no study had demonstrated cross-sensor validation. This finding reinforces the need for external validation and domain adaptation strategies.

Several limitations are common across the PPG-based apnea detection literature. First, many studies use fewer than 100 patients, although larger studies such as Jothi et al. [70] provide important exceptions. Second, evaluation is commonly performed on a single dataset without external validation. Third, Massie et al. [18] demonstrated central-versus-obstructive event classification and central-apnea screening from finger PPG, but evidence for direct three-class subject-level discrimination using a single-wavelength wearable ring remains limited. Fourth, continuous AHI estimation has received little attention, with Papini et al. [72] reporting a correlation of $r = 0.61$. Fifth, the reviewed studies provide little evidence on robustness to cardiac arrhythmias such as PVCs, which can alter PPG morphology and confound apnea-related patterns.

3.3 Transfer Learning in Biosignal Processing

Transfer learning [39] has been applied to biosignal processing to address the problem of limited labeled data in target domains. The general approach is to train a model on a large source dataset and then adapt it to a smaller target dataset, either by fine-tuning the entire model or by retraining only the final classification layers while keeping the learned representations frozen.

In sleep analysis, Radha et al. [20] demonstrated transfer learning from ECG to PPG for sleep staging. A model trained on a large ECG dataset (292 participants) was adapted to a smaller PPG dataset (60 participants) by retraining selected layers, achieving a Cohen’s kappa of 0.65 for 4-class staging compared to 0.56 without transfer. Li et al. [21] also transferred ECG-based sleep staging models to PPG using data from wrist-worn wearables. Both studies demonstrate cross-signal transfer learning in sleep analysis but address sleep staging rather than apnea detection. Vaquerizo-Villar et al. [85] applied CNNs with PPG for automatic sleep staging in children with sleep apnea. Birrer et al. [86] reviewed the reliability of wearable devices for sleep staging and noted an increasing trend toward combining accelerometer and PPG data.

In cardiac signal processing, domain adaptation techniques have been used to transfer arrhythmia detection models across different ECG databases. Deep CORAL [87] aligns the covariance of learned feature distributions between source and target domains. Batch normalization adaptation recalculates the running statistics of batch normalization layers using target-domain data without requiring labels. These unsupervised methods are attractive because they require no labeled target data, but their effectiveness depends on the assumption that the domain shift can be captured by simple distributional statistics. Jafari et al. [88] reviewed 239 papers on transfer learning in biomedical signals and found that EEG and ECG were the most studied modalities, with PPG receiving less attention. The review identified that supervised fine-tuning was the dominant approach, with unsupervised adaptation achieving lower performance.

The main gap in this area is the absence of transfer learning strategies for PPG-based apnea detection across different sensors and populations. The existing work transfers between signal modalities (ECG to PPG) or between tasks (sleep staging), but no study has addressed the problem of adapting a PPG-based apnea detection model trained on one sensor to work on a different sensor from a different manufacturer, population, or recording protocol. This gap is practically

important because different PPG sensors produce signals with different characteristics (sampling rates, dynamic ranges, filtering, noise profiles), and a model that only works on the training sensor has limited clinical value.

Cross-sensor transfer is arguably harder than cross-signal transfer (ECG to PPG) because the domain shift is more subtle and harder to characterize. When transferring from ECG to PPG, the source and target signals have fundamentally different morphology (QRS complexes vs. pulse waves), and the model must learn to extract the same clinical information from a different physical measurement. The domain shift is large but predictable. In contrast, when transferring between two PPG sensors, the signals have the same general morphology (pulse waves with systolic peaks and diastolic phases), but differ in ways that are difficult to enumerate: LED wavelength, photodetector sensitivity, analog filtering, analog-to-digital converter resolution, sampling timing jitter, optical coupling to the skin, and ambient light rejection. These differences are often undocumented by sensor manufacturers and cannot be corrected by simple preprocessing transformations. This work addresses this gap through the frozen backbone strategy described in Chapter 4.

3.4 Summary and Identified Gaps

Table 2 summarizes the main studies reviewed in this chapter, organized by task.

Table 2. Summary of representative studies in the three areas reviewed.

Study	Signal	Task	Key Result
Osathitporn 2023 [52]	PPG	RR estimation	MAE 1.66 (BIDMC), 1.59 (CapnoBase)
Bian 2020 [50]	PPG	RR estimation	MAE 2.5 ± 0.6 bpm (combined), ResNet
Pimentel 2017 [46]	PPG	RR estimation	MAE 1.9 (CapnoBase), quality-weighted fusion
Levy 2023 [67]	SpO ₂	Apnea detection	ICC 0.92–0.96 (AHI), N=12,923, OxiNet
Hayano 2020 [68]	PPG	Apnea detection	AUC 0.84, N=41, wrist watch
Jothi 2022 [70]	PPG	Apnea detection	Deep networks trained on N=1,375
Wei 2023 [73]	PPG	Apnea detection	Acc. 82.0/93.6%, MS-net
Radha 2021 [20]	ECG→PPG	Sleep staging TL	κ 0.65, N=60 target
Massie 2023 [18]	PPG	OSA/CSA analysis	Event typing and central-apnea screening

Taken together, these studies span different signals, cohort sizes, and evaluation metrics. The highest reported scores tend to come from SpO₂- and ECG-based methods, while PPG studies range from small single-site cohorts to larger exceptions such as the 1,375-subject study by Jothi et al. [70]. It should be stressed that the reported scores correspond to different targets and metrics (AUC, accuracy, F1, and ICC) and are therefore not directly comparable; they indicate the general landscape rather than a strict ranking.

Based on this review, four gaps are identified. First, many PPG-based apnea detection studies use small cohorts, and none of the reviewed raw-PPG studies was trained on more than 2,000 patients. Second, reviewed PPG-based methods are generally evaluated on the same dataset and sensor used for training, leaving degradation on different hardware insufficiently quantified. Third, although Massie et al. [18] established that finger PPG can support central-versus-obstructive event analysis, direct three-class subject-level discrimination between Normal, OSA, and CSA using single-wavelength wearable-ring PPG remains insufficiently studied. Fourth, no reviewed study developed and evaluated a transfer strategy specifically for adapting a PPG-based apnea detector across sensors and populations.

This work addresses these gaps through the CoTELMo architecture trained on 2,056 MESA patients (Gap 1), evaluation on CFS, additional NSRR cohorts, and Absolute Rest with different sensors (Gap 2), an exploratory three-class phenotype analysis of Normal, OSA, and CSA on wearable-ring PPG (Gap 3), and a frozen-backbone adaptation strategy with quantified recovery rates (Gap 4).

3.5 Key Remarks

This chapter reviewed the literature on respiratory rate estimation from PPG, sleep apnea detection, and transfer learning in biosignal processing. Respiratory-rate methods report competitive errors on BIDMC and CapnoBase, but they have not been connected to apnea detection. Apnea detection has been studied with ECG, SpO₂, and PPG; PPG-based methods show promise, but cross-sensor evaluation remains uncommon and phenotype evidence is limited to specific finger-based systems. Transfer learning has been applied to sleep staging across signal modalities, but not specifically to PPG-based apnea detection across sensors. These gaps define the research space addressed by the following chapter.

4 Materials and Methods

The experimental framework rests on five primary datasets spanning different populations, sensors, and clinical settings, each serving a distinct role in the pipeline. The sections below describe these datasets and the methods developed to address each research question, organized in the sequence in which they were applied: signal preprocessing, classical baseline, deep learning architectures, continuous AHI estimation, transfer learning for cross-sensor adaptation, and phenotype classification.

4.1 Datasets

Five primary datasets were used in this work, each serving a distinct role in the pipeline. Two short-recording clinical datasets (BIDMC and CapnoBase) were used for respiratory rate regression. One large-scale epidemiological dataset (MESA) provided the training data for the apnea detection model. One independent sleep study (CFS) was used for external validation and domain adaptation evaluation. One commercial wearable dataset (Absolute Rest) served as the final clinical validation on the target device. In addition, three further NSRR cohorts (HeartBEAT, BestAIR, and HomePAP) were used solely to characterize sensor-level domain shift and signal quality (Section 4.2); they were not used for model training or adaptation and are therefore not included in the primary dataset summary. The following sections describe each primary dataset in detail.

The Beth Israel Deaconess Medical Center (BIDMC) dataset was compiled by Pimentel et al. [46] and is publicly available through PhysioNet [89]. It contains recordings from 53 patients in the intensive care unit (ICU), each approximately 8 minutes long. The signals include PPG, ECG, and a reference respiratory rate derived from impedance pneumography. The native sampling rate is 125 Hz. The ICU setting introduces clinically relevant noise sources: patients are often sedated, mechanically ventilated, or receiving vasoactive medications that alter cardiovascular dynamics and peripheral perfusion. These conditions produce PPG signals that are more challenging to process than recordings from healthy volunteers, making BIDMC a demanding benchmark for respiratory rate estimation. In this work, the BIDMC dataset was split at the patient level into 45 subjects for training and validation, and 8 subjects for held-out testing, yielding 1,390 valid training windows and 248 test windows. This dataset was used exclusively for training and evaluating the respiratory rate regression model described in Section 4.4.

The CapnoBase dataset was introduced by Karlen et al. [45] and contains recordings from 42 subjects undergoing elective surgery or routine anesthesia, each approximately 8 minutes long. The signals include PPG and capnography (end-tidal CO_2), with the capnography signal providing a reference respiratory rate. Capnography measures the CO_2 concentration in exhaled air and provides an accurate non-invasive reference for respiratory rate, as it directly detects each breath regardless of the breathing pattern. The native sampling rate is 300 Hz, which was resampled to 125 Hz to match the BIDMC preprocessing pipeline. The data was split into 29 development subjects and 13 fixed test subjects. The development partition was combined with the BIDMC development partition, whereas the 13 test subjects were excluded from every training fold and used as a held-out target-domain evaluation. Because CapnoBase contributed development data, this final result is not a strict dataset-level zero-shot external validation.

The Multi-Ethnic Study of Atherosclerosis (MESA) Sleep Study is a large-scale epidemiological cohort available through the National Sleep Research Resource (NSRR) [90, 91]. It contains full-night PSG recordings from 2,237 community-dwelling adults aged 54 to 95 years (mean 69.6), with 53.6% female participants. The study population is ethnically diverse, including Caucasian, African American, Hispanic, and Chinese American participants, which provides demographic variation that is absent from most single-center clinical studies. Each recording includes multiple channels, of which the PPG signal (labeled “Pleth” in the EDF files) was sampled at 256 Hz using a Nonin 8000 series pulse oximeter, a clinical-grade finger sensor that produces high-quality PPG with consistent morphology. Respiratory event annotations were provided in NSRR XML format, scored according to AASM criteria [28], and include obstructive apneas, central apneas, mixed apneas, and hypopneas with their precise start times and durations. These event-level annotations are critical because they enable the strong labeling strategy described in Section 4.2, which assigns per-window labels based on actual event overlap rather than relying solely on patient-level AHI thresholds.

After excluding participants with missing AHI values or missing respiratory annotations, 2,056 patients remained for analysis. The AHI distribution across the four standard severity categories is presented in Table 3. The dataset has a roughly uniform distribution across severity levels, with no single category dominating, which is favorable for training classification models.

Table 3. AHI severity distribution in the MESA dataset ($N = 2,056$).

Severity	AHI Range	N (%)
Normal	< 5	414 (20.1%)
Mild	5–15	643 (31.3%)
Moderate	15–30	518 (25.2%)
Severe	≥ 30	481 (23.4%)

The MESA dataset served as the primary training set for the CoTELMo apnea detection model. Its combination of large sample size, event-level annotations, and diverse severity distribution made it well suited for this purpose.

The Cleveland Family Study (CFS) is a longitudinal family-based cohort available through the NSRR [91, 92]. It was originally designed to study the familial aggregation of sleep apnea, and its participants were recruited from families in which at least one member had been diagnosed with OSA. This recruitment strategy means that the cohort includes a higher proportion of genetically predisposed individuals than population-based studies like MESA, and the age distribution is broader (including children and young adults). The PSG recordings include a PPG channel labeled “PlethWV” sampled at 128 Hz, half the rate of MESA (256 Hz), which means that the CFS PPG signal captures frequencies up to 64 Hz rather than the 128 Hz available in MESA. Although both rates far exceed the clinically relevant band for respiratory analysis (below 1 Hz), the different sampling rates contribute to subtle differences in the digital noise floor and anti-aliasing filter characteristics that may affect the learned CNN features. After signal quality filtering (described in Section 4.2), 324 patients were retained for analysis. The AHI distribution is summarized in Table 4. Unlike MESA, this dataset has a larger proportion of normal subjects (54.6%), providing a different class balance for evaluating model generalization.

Table 4. AHI severity distribution in the CFS dataset ($N = 324$).

Severity	AHI Range	N (%)
Normal	< 5	177 (54.6%)
Mild	5–15	64 (19.8%)
Moderate	15–30	43 (13.3%)
Severe	≥ 30	40 (12.3%)

The CFS was selected for external validation and domain adaptation evaluation because it shares several characteristics with MESA (both are NSRR datasets with scored respiratory events) while differing in the PPG sensor hardware and patient population, creating a controlled setting for studying domain shift.

The Absolute Rest (AR) dataset consists of overnight PPG recordings collected using the Absolute Rest smart ring, a commercial wearable device designed for sleep monitoring in elite athletes. The data was provided through a research collaboration with Absolute Rest, a US-based sleep technology startup. A total of 54 recording sessions were collected from multiple users, of which 52 had valid data files (two Normal recordings were excluded for missing files). Recordings were stored as individual Parquet files containing the PPG waveform (“Pleth” column). Each recording spanned a full night of sleep, with typical durations of 7 to 8 hours.

The ring sensor has a variable native sampling rate that differs across sessions. The most common rate was 71.4 Hz (22 sessions), followed by 125.0 Hz (15 sessions), with a few sessions at 83.3, 76.9, or 40.0 Hz. All recordings were resampled to a uniform rate during preprocessing (125 Hz for respiratory rate analysis, 64 Hz for CoTELMo inference).

The recordings were acquired using a proprietary ring-form-factor PPG sensor worn on the finger during full-night sleep sessions. Unlike the clinical PSG datasets (MESA and CFS), which use clip-on pulse oximeters with standardized placement and controlled recording environments, the Absolute Rest ring is a consumer-grade wearable device that participants wore in their habitual sleep environment without technician supervision. This distinction is important because it introduces additional sources of variability: finger diameter differences affect optical coupling, hand positioning during sleep introduces intermittent motion artifacts, and ambient temperature variations in uncontrolled bedroom environments modulate peripheral perfusion. The dataset therefore represents a realistic deployment scenario for wearable sleep apnea screening rather than a controlled laboratory setting.

The recordings were labeled with clinical diagnoses into four conditions: Normal breathing (34 sessions), Severe OSA with $AHI \geq 30$ (6 sessions), CSA (6 sessions), and Premature Ventricular Contractions or PVC (6 sessions). The PVC recordings were held out from all training and

used exclusively as a robustness check, since cardiac arrhythmias can produce PPG morphological changes that might be confused with respiratory abnormalities. After excluding the PVC group, 46 sessions were used for the main classification experiments (34 Normal, 6 Severe OSA, 6 CSA). This dataset represents the intended deployment scenario for the pipeline, serving as the final clinical validation on the target hardware.

Table 5 provides an overview of the five primary datasets and their roles in this work.

Table 5. Summary of the primary datasets used in this work.

Dataset	N	PPG (Hz)	Duration	Role
BIDMC	53	125	~8 min	RR training
CapnoBase	42	300	~8 min	RR held-out test
MESA	2,056	256	Full night	Apnea model training
CFS	324	128	Full night	External validation
Absolute Rest	46 + 6	Variable	Full night	Clinical validation

Figure 8 compares the severity distributions across the three main datasets, illustrating the different class balance challenges faced during training and validation.

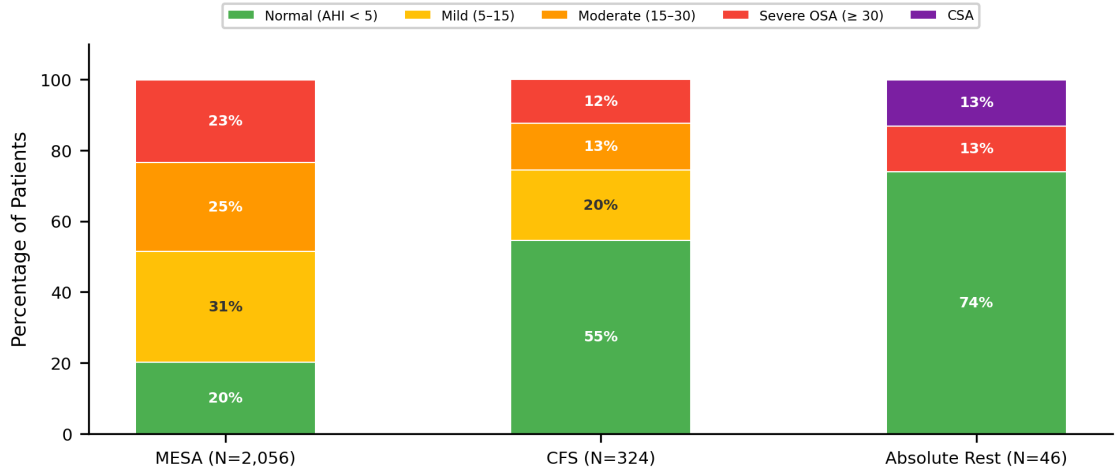


Fig. 8. AHI severity distribution across the three main datasets. MESA has a roughly uniform distribution across severity levels, CFS is dominated by normal subjects (55%), and, after excluding the PVC robustness group, Absolute Rest contains only Normal, Severe OSA, and CSA classes. The different class balances require dataset-specific evaluation strategies.

4.2 Signal Preprocessing

Two preprocessing pipelines served the two main tasks: the respiratory rate pipeline operated at 125 Hz on short clinical recordings, and the apnea detection pipeline at 64 Hz on full-night recordings. Both used 30-second windows, the unit chosen to match the American Academy of Sleep Medicine (AASM) respiratory-event scoring convention [28] and the common PSG sleep-epoch duration. Since a scored apnea event lasts at least 10 seconds, a 30-second window captures at least one full event with its surrounding breathing; events crossing a window boundary are recovered through the 5-second overlap of the labeling strategy described below. Shorter windows (10 seconds) would raise temporal resolution but fragment longer events across windows, complicating labeling; longer windows (60 seconds) would add context but reduce training samples per patient and dilute short-event signals within a larger normal-breathing context. The pipelines differed in overlap, normalization, quality filtering, and extracted features, summarized in Figure 9.

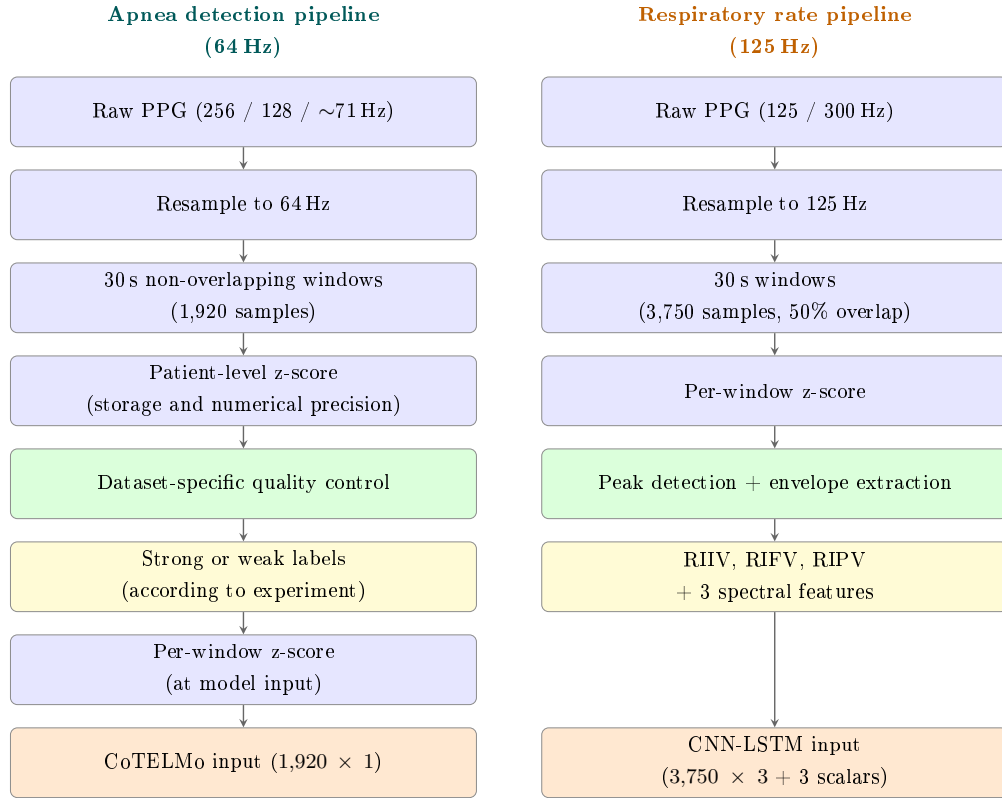


Fig. 9. Signal preprocessing pipelines. The apnea pipeline uses patient-level normalization for storage, dataset-specific quality control, and per-window normalization at model input. The respiratory-rate pipeline uses 50%-overlapping windows at 125 Hz and extracts RIIV, RIFV, RIPV, and three spectral features.

For the respiratory rate regression task (BIDMC and CapnoBase datasets), the PPG signals were first bandpass filtered between 0.5 and 8.0 Hz to isolate the cardiac pulse component and its harmonics while removing baseline drift and high-frequency noise, then segmented into 30-second windows at the native or resampled rate of 125 Hz, producing 3,750 samples per window. A 50% overlap was used between consecutive windows to increase the number of training samples. Each signal was normalized using z-score standardization (zero mean, unit variance) computed per signal channel within each window.

Three respiratory modulation signals were extracted from each PPG window using peak detection and envelope reconstruction. Peaks were detected using the `scipy.signal.find_peaks` function with a minimum inter-peak distance of 0.4 seconds (50 samples at 125 Hz) and a minimum height threshold of 0.5 standard deviations above the mean. From the detected peaks and valleys, three respiratory-induced variation signals were computed: RIIV, obtained from the baseline wander of the PPG signal through cubic spline interpolation of the peak positions [12]; RIFV, derived from the instantaneous heart rate as the reciprocal of the inter-peak intervals; and RIPV, computed from the temporal width of each cardiac pulse. Each envelope signal was lowpass filtered using a fourth-order Butterworth filter with a cutoff frequency of 0.5 Hz and resampled to match the original window length through cubic spline interpolation [14]. A systematic correlation analysis confirmed that RIIV captures the strongest respiratory modulation from the PPG signal, with a mean Pearson correlation of 0.450 against the reference impedance pneumography waveform, compared to 0.193 for the raw PPG amplitude envelope.

For the apnea detection task, full-night PPG recordings from MESA, CFS, and Absolute Rest were preprocessed using a common pipeline. All signals were first resampled from their native sampling rates to a target rate of 64 Hz using Fourier-based resampling (`scipy.signal.resample`). The choice of 64 Hz, rather than the 125 Hz used for respiratory rate extraction, was motivated by both physiological and practical considerations. Physiologically, apnea detection relies on slower signal modulations (respiratory effort and autonomic variations) rather than the precise peak-to-peak temporal resolution required for interval estimation. By the Nyquist theorem, a 64 Hz sampling rate captures frequencies up to 32 Hz, encompassing all clinically relevant cardiac and respiratory bands. Practically, this reduction reduced the computational cost during training on the 2.4 million windows of the MESA dataset.

The resampled signals were segmented into non-overlapping 30-second windows of 1,920 samples each. Non-overlapping windows were used for apnea detection, unlike the 50% overlap used for respiratory rate, because the full-night recordings already provide sufficient training data without augmentation: a typical 7-hour recording yields about 840 windows per patient, and the MESA dataset contains over 2.4 million windows in total. Overlap would approximately double the number of windows while introducing substantial redundancy between adjacent segments, increasing computational cost without substantially increasing diversity.

Normalization was applied in two stages. First, a per-patient z-score was computed across the entire recording before windowing, reducing inter-subject amplitude variability and improving compatibility with float16 storage. Second, within the CoTELMo training pipeline, each 30-second window was independently standardized to zero mean and unit variance before being fed to the CNN, providing stable input statistics for the convolutional front-end.

Quality filtering was applied at the window level after per-patient normalization. Windows with a standard deviation below 10^{-6} , evaluated on the already-normalized signal, were discarded as flat-line artifacts. Windows containing more than 10% missing values (NaN) were also discarded; any remaining NaN values in valid windows were set to zero, which corresponds to the patient mean in the normalized space. A minimum threshold of 10 valid windows per patient was required for inclusion in the analysis. These criteria were deliberately conservative for MESA: a preliminary analysis showed that signal quality correlates negatively with AHI severity (Pearson $r = -0.174$, $p < 10^{-15}$), suggesting that stricter filtering could disproportionately exclude patients with more severe apnea and introduce systematic bias.

For external validation datasets (CFS, HeartBEAT, BestAIR, HomePAP), a more structured Signal Quality Index (SQI) was computed per window using four metrics inspired by the approach of Elgendi [93] and Orphanidou et al. [47]. The amplitude SQI required a minimum peak-to-peak amplitude of 0.3 standard deviations, evaluated on the per-window z-scored signal. The flatline SQI rejected windows where more than 20% of consecutive samples had identical values. The template matching SQI computed the correlation between each cardiac cycle and a patient-specific template, requiring a minimum correlation of 0.5. The skewness SQI required the window skewness to fall within the range $[-3.0, 1.0]$, consistent with normal PPG morphology. A window was considered

valid if the amplitude and flatline criteria were both satisfied, and at least one of the template or skewness criteria was met. A patient was included in the analysis if at least 50% of their windows were valid.

The pass rates varied across datasets: CFS retained 324 of 324 patients (100%, mean SQI 98.8%), BestAIR retained 129 of 134 (96.3%, mean SQI 70.7%), HeartBEAT retained 251 of 309 (81.2%, mean SQI 67.3%), and HomePAP retained only 56 of 96 (58.3%, mean SQI 56.1%). These differences reflect variations in sensor hardware, electrode placement, and recording conditions across studies. Figure 10 provides a visual summary of the quality filtering pipeline and its outcomes.

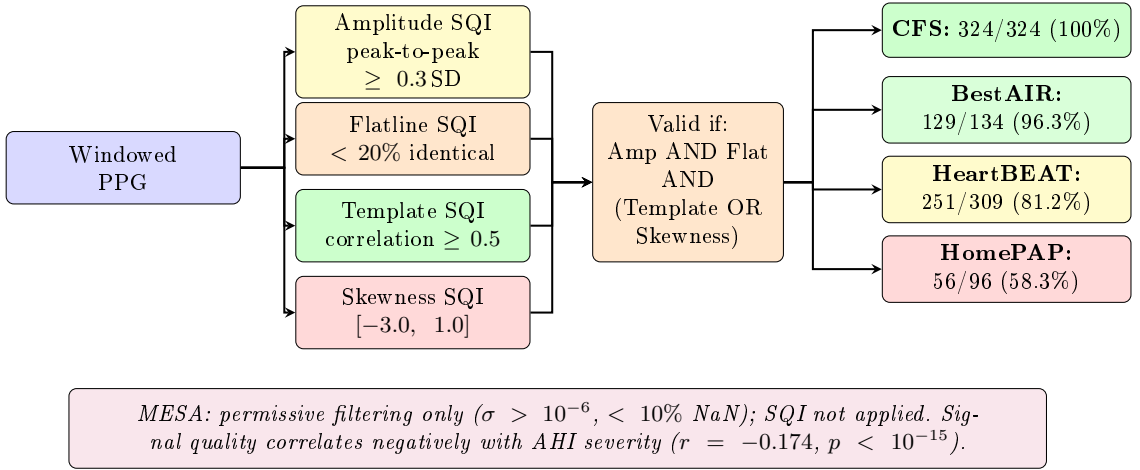


Fig. 10. Signal quality index filtering pipeline for external validation datasets. Four quality criteria (amplitude, flatline, template matching, and skewness) are combined using a logical rule (amplitude AND flatline AND (template OR skewness)). Pass rates vary from 100% (CFS) to 58.3% (HomePAP). MESA uses permissive filtering only, as signal quality correlates negatively with AHI severity.

A controlled experiment was conducted to determine whether bandpass filtering the raw PPG signal before model input improved classification performance. A fourth-order Butterworth bandpass filter with a passband of 0.5–8 Hz was compared with no filtering in an A/B test using the classical machine learning baseline on MESA. The bandpass filter reduced AUC from 0.796 to 0.795, a difference of 0.001. This result is consistent with the view that the CoTELMo convolutional front-end can learn frequency-selective representations directly from raw PPG, making an explicit bandpass filter unnecessary in this setup. Accordingly, the apnea detection pipeline does not apply bandpass filtering to the raw PPG input; the signal is resampled to 64 Hz and z-score normalized at the patient level before windowing, with no additional spectral pre-processing.

Two labeling strategies were employed in this work, and the distinction between them proved critical to model performance. Weak labels were derived directly from the patient-level AHI values provided in the MESA dataset and required only a patient-level thresholding step. Strong labels were constructed from the respiratory event annotations parsed from the MESA XML files, using a custom labeling function that assigned window-level supervision based on event overlap.

Weak labels assign a single binary label to each patient based on their overall AHI value. For example, at the $\text{AHI} \geq 15$ threshold, a patient is labeled as positive if their AHI is 15 or above, and negative otherwise. This label is then applied uniformly to all windows from that patient. Weak labels are simple to obtain but introduce substantial noise at the window level, because individual 30-second windows from a severe apnea patient may contain perfectly normal breathing.

Strong labels assign a binary label to each individual 30-second window based on the respiratory event annotations available in the MESA dataset. A window was labeled as positive if any scored respiratory event overlapped with the window by at least 5 seconds. This threshold corresponds to half the minimum event duration defined by the AASM scoring criteria (10 seconds) [28] and reduces the risk of missing events near window boundaries. A stricter threshold, such as 15 seconds, would miss events that straddle adjacent windows. Using the MESA event-level annotations, approximately 2,452,790 windows were labeled, of which 54.9% were positive. This strong labeling strategy is only possible when event-level annotations are available, which limits it to the MESA training set. External validation and fine-tuning on CFS and Absolute Rest relied on weak labels because event-level annotations were either unavailable or impractical to obtain.

The correlation between the per-patient apnea ratio (fraction of windows labeled positive under the strong labeling scheme) and the clinical AHI was $\rho = 0.854$, supporting the consistency between the window-level labels and the patient-level diagnosis. Figure 11 illustrates the difference between weak and strong labeling for a single patient.

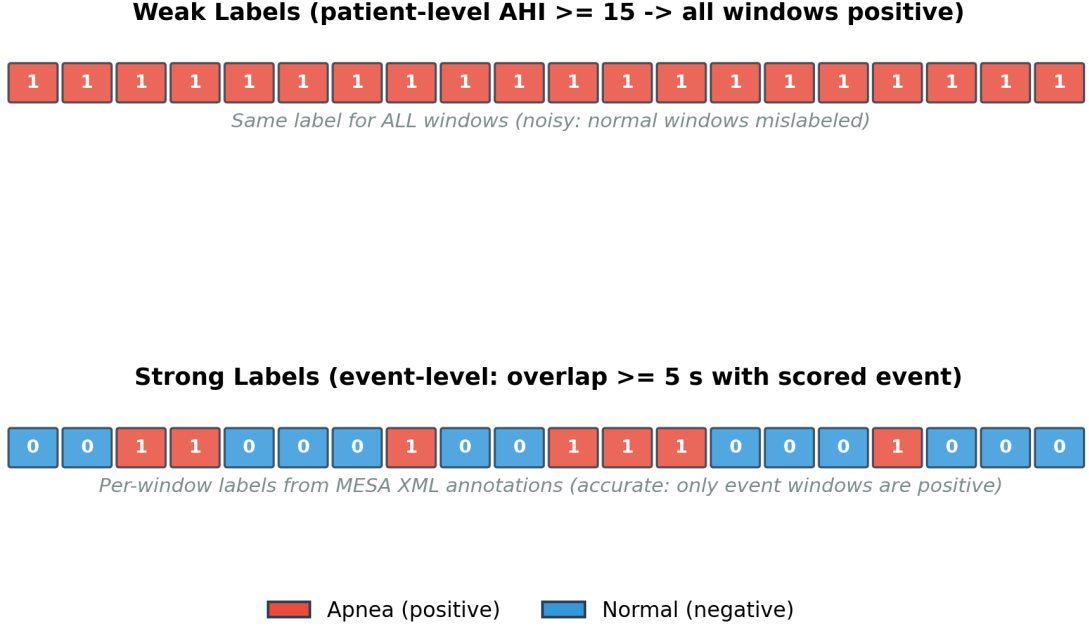


Fig. 11. Weak vs. strong labeling strategies for a single patient recording ($\text{AHI} \geq 15$). Top: weak labels assign the same positive label to all 30-second windows regardless of their actual content, introducing substantial noise from normal-breathing windows. Bottom: strong labels use event-level annotations from the MESA XML files, assigning positive labels only to windows that overlap with a scored respiratory event by at least 5 seconds.

4.3 Classical Machine Learning Baseline

Before developing deep learning models, a classical machine learning baseline was established on the MESA dataset using hand-crafted features. This baseline served two purposes: providing a performance floor against which deep learning improvements could be measured, and identifying which classical approaches could generalize across datasets.

A total of 430 features were extracted from each patient’s PPG recording, organized into four categories. Cardiac features included heart rate statistics (mean, standard deviation, percentiles), inter-beat interval variability, and pulse morphology measures. Respiratory features included the three modulation signals (RIIV, RIFV, RIPV) and their spectral characteristics. Statistical features captured the distribution of window-level signal properties across the full recording. Morphological features described the shape of individual PPG pulses, including pulse width, rise time, and peak-to-valley ratios. In a second phase, 301 additional advanced features were computed, including perfusion index, an ODI-surrogate derived from PPG amplitude patterns, nonlinear pulse rate variability measures, and cardiopulmonary coupling estimates.

Four classifiers were compared: Logistic Regression (LR), Random Forest (RF), ExtraTrees [94], and Gradient Boosting Machine (GBM). Each was evaluated on three tasks. The first two were internal classifications on MESA using 5-fold cross-validation: $\text{AHI} \geq 5$ (any apnea, including mild cases) and $\text{AHI} \geq 15$ (the clinical treatment threshold for moderate-to-severe OSA). $\text{AHI} \geq 5$ was included to assess whether the features could detect milder cases, where the classification problem is more challenging; $\text{AHI} \geq 15$ served as the primary clinical reference for the main reported comparisons. The third task was cross-dataset transfer, where models trained on MESA were applied directly to the Absolute Rest data at $\text{AHI} \geq 15$.

The tree-based models (RF, GBM, ET) achieved near-zero specificity on the internal task at $\text{AHI} \geq 5$, classifying virtually every patient as apnea-positive (specificity 4.5–12.1%). Only LR produced balanced predictions (approximately 56% sensitivity and 56% specificity). In the cross-dataset validation, LR maintained an AUC of 0.675 when transferred to the Absolute Rest data, while GBM collapsed to 0.435, which is worse than random. This pattern reflects a classical bias-variance tradeoff under domain shift: high-complexity models overfit to dataset-specific feature distributions and fail on new data, while LR, constrained by its linear decision boundary, captures more generalizable patterns. Based on these results, LR was selected as the baseline classifier for all subsequent comparisons.

Nested cross-validation was performed using Optuna [95] to optimize the LR regularization parameters. The outer loop used 5-fold stratified cross-validation, and the inner loop used 3-fold cross-validation for hyperparameter selection. This nested design ensures that the reported AUC values are unbiased estimates of generalization performance. The optimization ran for approximately 9 hours. In the optimized PPG-only run, LR reached an AUC of 0.796 at $\text{AHI} \geq 15$ and 0.742 at $\text{AHI} \geq 5$ using 183 selected features from the 430-feature set, consistent with the expectation that milder cases are harder to classify from PPG alone. With the 301 additional advanced features (731 total), performance at $\text{AHI} \geq 15$ improved slightly to 0.807, indicating diminishing returns from further feature engineering.

As a matched signal comparison, separate LR classifiers were evaluated using PPG-only, SpO_2 -only, and combined PPG+ SpO_2 feature sets under the same 5-fold cross-validation protocol. The candidate feature sets contained 430 PPG features, 54 SpO_2 features, and 484 combined features.

The SpO_2 features were grouped into desaturation indices, time-below-threshold metrics, statistical descriptors, variability metrics, and spectral features capturing cyclic desaturation in the 0.01–0.04 Hz band. This matched comparison keeps the evaluation protocol constant across sensors, while the optimized PPG-only run reported above remains a separate feature-selected baseline.

In this matched comparison, the SpO_2 -only classifier achieved an AUC of 0.953 at $\text{AHI} \geq 15$, 19 percentage points above the PPG-only baseline of 0.765. This result is expected and consistent with the literature on oximetry-based screening, because the AHI definition used in MESA counts events that include oxygen desaturation of at least 3% as a criterion. Features such as ODI3 and CT90 therefore capture information that is structurally linked to the outcome variable. The SpO_2 result does not represent a fair competitor to PPG-based models; it serves as a performance ceiling, quantifying how much information is available when the sensor has direct access to the physiological consequence of apnea rather than its upstream cardiovascular correlates. Narrowing this gap was one of the motivating objectives for the deep learning approach.

4.4 Respiratory Rate Regression

The first deep learning component of the pipeline addresses respiratory rate estimation from PPG. This task served as an initial validation that the PPG signal contains recoverable respiratory information, and the features and architectural decisions explored here informed the subsequent apnea detection work. The respiratory rate regression problem differs from apnea detection in several ways: the input is multi-channel (three PPG-derived respiratory modulation signals and spectral features), the output is continuous rather than binary, and the recordings are short (8 minutes) rather than full-night. These differences motivated a separate architecture from the CoTELMo model, although the underlying CNN-LSTM hybrid design is shared between both tasks.

An initial approach attempted direct respiratory waveform reconstruction using a 1D U-Net architecture with a Pearson correlation loss. The model received the PPG signal as input and was trained to reconstruct the reference impedance pneumography waveform sample-by-sample. Three input variants were tested: the raw PPG, the Hilbert-transformed amplitude envelope, and a cross-correlation aligned version. All three variants failed, with the best achieving a test correlation of -0.027 , no better than random. This result confirmed that the respiratory modulation embedded in the raw PPG waveform is too weak for direct signal-to-signal mapping, and motivated the shift to

respiratory rate estimation using extracted modulation signals (RIIV, RIFV, RIPv) that amplify the respiratory component through peak detection and envelope reconstruction.

Six features were used as input to the final respiratory rate model, divided into two groups. The temporal group consisted of three respiratory-induced variation signals (RIIV, RIFV, and RIPv), each represented as a time series of 3,750 samples per 30-second window at 125 Hz. These three signals were stacked as channels, forming a $3,750 \times 3$ input tensor.

The spectral group consisted of three scalar features computed from the frequency domain: the dominant frequency of the RIIV signal (`riiv_df`), the dominant frequency of the RIPv signal (`ripv_df`), and the power ratio of the RIFV signal (`rifv_pr`), defined as the ratio of spectral power in the respiratory band (0.1–0.5 Hz) to total spectral power. These three scalars captured global frequency-domain characteristics that complement the temporal patterns.

An important design decision was the exclusion of ECG-derived respiration (EDR). The baseline model used RIIV and EDR as its two input channels, achieving the lowest MAE on BIDMC (1.37 bpm). However, EDR requires an ECG signal, which is not available on wrist or ring-based wearable devices. The final model replaced the ECG dependency with RIFV and RIPv, two additional PPG-derived signals, making the approach fully PPG-only at the cost of a small increase in MAE (from 1.37 to 1.70 bpm on BIDMC). Although four respiratory modulation mechanisms are described in the literature (Chapter 2), the fourth signal, RIAV (derived from the PPG valley amplitudes), was excluded from the final model after a feature ablation study showed that its contribution was marginal compared to the other three signals.

Beyond the hardware constraint, a second motivation for the final architecture was cross-dataset generalization. When the sequential single-scale model was evaluated on CapnoBase without re-training, performance degraded substantially relative to its BIDMC result, revealing a domain gap between ICU recordings and elective-surgery recordings acquired with different monitoring equipment. Increasing the number of channels within a sequential design did not resolve this gap. In the parallel multi-scale architecture, independent branches capture respiratory patterns at short, medium, and long temporal scales before fusion. Training this architecture on a development super-dataset that merged BIDMC and CapnoBase with sample weighting to compensate for the size

imbalance addressed both issues simultaneously. The modest reduction in BIDMC performance (1.37 to 1.70 bpm) that resulted from removing the ECG dependency was offset by improved performance on the held-out CapnoBase test partition.

The final respiratory rate model uses a multi-scale CNN-LSTM architecture with a dual-input design that processes temporal and spectral features in parallel branches before fusion. Figure 12 illustrates the architecture.

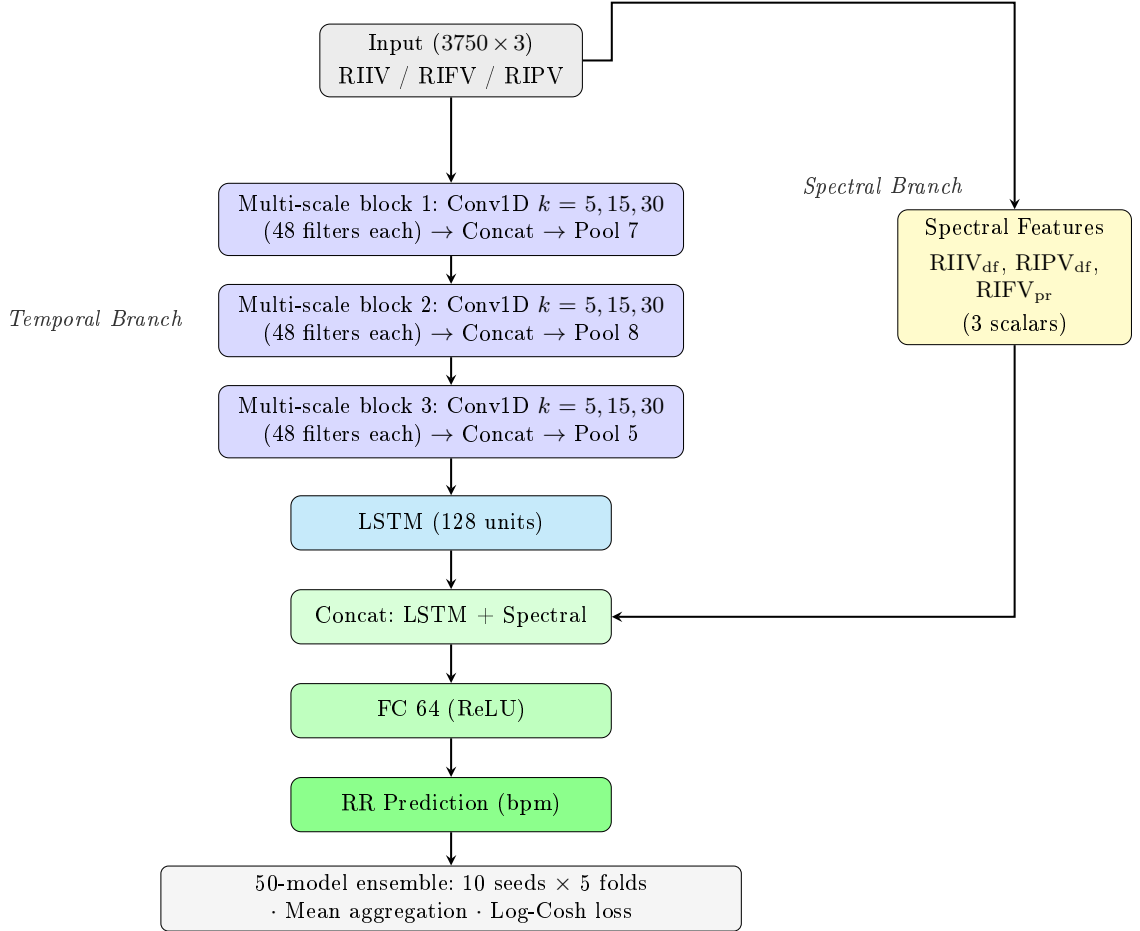


Fig. 12. Multi-scale CNN-LSTM architecture for respiratory rate estimation. The temporal branch contains three stacked blocks; within each block, parallel convolutions at kernel sizes 5, 15, and 30 are concatenated before pooling (7, 8, and 5 across successive blocks). The resulting sequence is processed by an LSTM and fused with three scalar spectral features.

The temporal branch consists of three multi-scale convolutional blocks. Each block applies three parallel 1D convolutions with different kernel sizes (5, 15, and 30 samples) and 48 filters each, capturing respiratory patterns at different temporal scales. The outputs of the three parallel convolutions are concatenated along the feature dimension, producing 144 features per timestep.

This is followed by max pooling and dropout (rate 0.2). The three blocks use pool sizes of 7, 8, and 5, progressively reducing the temporal resolution. After the three convolutional blocks, an LSTM layer with 128 hidden units processes the sequence, and its final hidden state is passed through a dropout layer (rate 0.1).

The spectral branch takes the three scalar features as direct input without any transformation. The outputs of both branches are concatenated into a single vector, which is passed through a fully connected layer with 64 units (ReLU activation, dropout 0.1) and a final linear output layer that produces the predicted respiratory rate in breaths per minute.

The model was trained on a development super-dataset formed by merging 45 BIDMC and 29 CapnoBase subjects, with a weighting factor of 1.7 applied to CapnoBase samples to compensate for its smaller size. GroupKFold cross-validation with 5 folds was used within this development set, with patient-level grouping to prevent leakage. Training was repeated across 10 random seeds, producing 50 models total (10 seeds $\times$ 5 folds). The 8 BIDMC and 13 CapnoBase test subjects were excluded from every fold; their final predictions were obtained by averaging all 50 model outputs. Figure 13 illustrates this strategy.

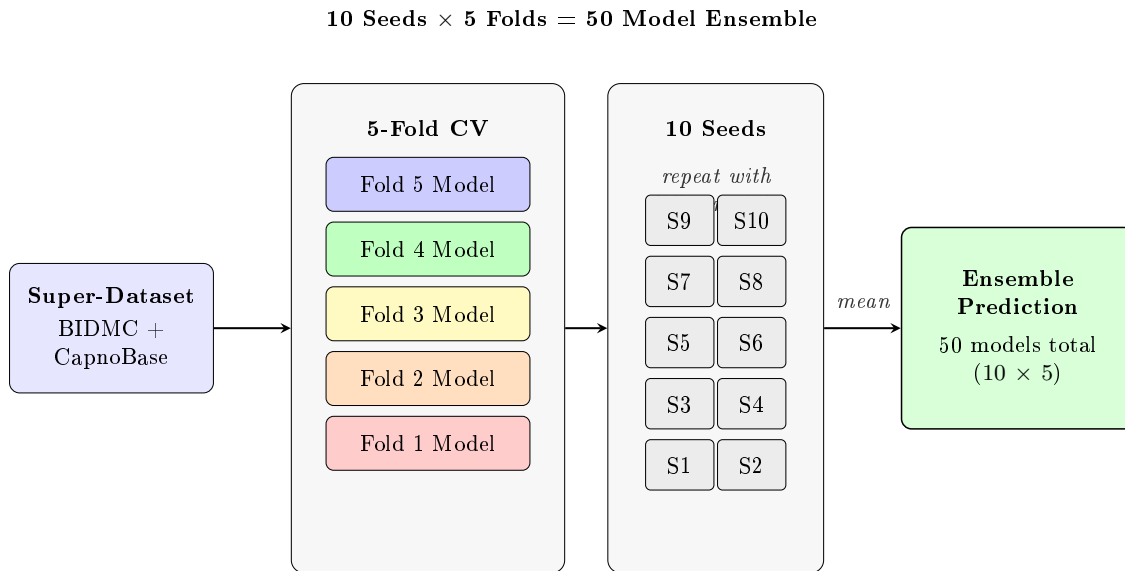


Fig. 13. Ensemble strategy for respiratory rate estimation. The combined BIDMC–CapnoBase development set is split into 5 patient-level folds, and training is repeated with 10 random seeds, producing 50 models. All models remain blind to the fixed BIDMC and CapnoBase test subjects, whose final prediction is the mean of the 50 outputs.

Three loss functions were compared: Mean Squared Error (MSE), Huber loss ($\delta = 1.0$), and Log-Cosh loss. The optimizer was Adam [96] with a learning rate of 2.71×10^{-4} and a batch size of 32. These hyperparameters, along with the architectural parameters (48 filters per scale, 128 LSTM units, dropout rates), were jointly optimized using Optuna [95] with the TPE sampler on the combined super-dataset, selecting the configuration that minimized the cross-validated MAE. Data augmentation was applied during training using additive Gaussian noise (5% of signal standard deviation) and random amplitude scaling (factor 0.8–1.2), both of which simulate the signal quality variations encountered in real-world recordings. Training ran for a maximum of 150 epochs with early stopping based on validation loss (patience of 20 epochs, restoring best weights). The model was implemented in TensorFlow/Keras on the NVIDIA GeForce RTX 3060 GPU.

A temporal consistency smoothing step was applied as post-processing to the ensemble predictions. A moving average filter with a window size of 3 consecutive predictions was used to attenuate prediction jitter between adjacent overlapping windows. The window size was selected through a grid search over odd values from 1 to 9, with 3 producing the lowest CapnoBase MAE.

4.5 Sleep Apnea Detection with CoTELMo

The central contribution of this work is the CoTELMo architecture for sleep apnea detection from raw PPG signals. The development of this model followed a progression from a simple CNN baseline trained with weak labels, through the introduction of event-level supervision and the choice of a clinically aligned aggregation strategy, culminating in the hybrid architecture that achieved the best results.

A 1D CNN with four convolutional blocks (32, 64, 128, and 256 filters) followed by global average pooling and a fully connected classifier was used as the initial baseline. When trained with weak labels (patient-level AHI threshold), this model achieved a patient-level AUC of only 0.621, barely above random chance. The poor performance was traced to the weak labeling strategy: because every window from a positive patient receives a positive label regardless of its actual content, the model receives contradictory supervision from the many windows in a severe apnea patient’s recording that contain normal breathing.

Switching to strong labels (event-level annotations from the MESA XML files) raised the window-level AUC to 0.771 and the patient-level AUC under mean aggregation to 0.823. This confirmed that label quality was the main bottleneck. The patient-level gain came from the change in supervision: weak labels assign the same patient-wide AHI category to every window, while strong labels mark only the windows that overlap with annotated events.

Converting window-level predictions to a patient-level score requires an aggregation strategy. Mean aggregation of all window probabilities produced a patient-level AUC of 0.823 under strong labels, an improvement of 0.20 over the weak-label baseline. The choice of aggregation strategy can further refine this result, because the patient signal of interest is concentrated in the minority of windows that contain apnea events. Averaging across all windows treats them as equally informative, while a top-K aggregation focuses on the windows that carry the diagnostic information.

An oracle analysis was conducted to establish the theoretical ceiling: using the ground truth window labels to compute the optimal patient score yielded a patient-level AUC of 0.928. This ceiling defined the maximum performance achievable with any aggregation strategy, given the underlying window-level classification quality.

Multiple aggregation strategies were tested, including mean, median, various percentiles, top-K percentages, and threshold-based counting. The top-K mean strategy averages only the K% highest window probabilities, enriching the patient score for the minority of windows most likely to contain respiratory events. Using the top 10% of window scores, the patient-level AUC rose from 0.823 to 0.865, a further gain of 0.04 over mean aggregation that recovered 93% of the oracle ceiling using CNN scores alone.

The full comparison across the 21 tested strategies shows that this advantage is not specific to a single value of K. The four best strategies were all upper-tail statistics: top-10% mean (0.865), top-5% mean (0.863), top-20% mean (0.862), and the 90th percentile (0.859), separated by less than 0.007 AUC. Statistics of central tendency ranked clearly below them (mean, 0.823; median, 0.798), and counting the windows above a fixed probability was the weakest family (0.793 to 0.839). The gain therefore comes from restricting the patient score to the upper tail of the window distribution, not from any particular choice of K. At its operating point, top-10% aggregation

also raised sensitivity from 0.730 to 0.815 at comparable specificity (0.774 to 0.778), which is the direction that matters for a screening tool. Figure 14 illustrates the strategy.

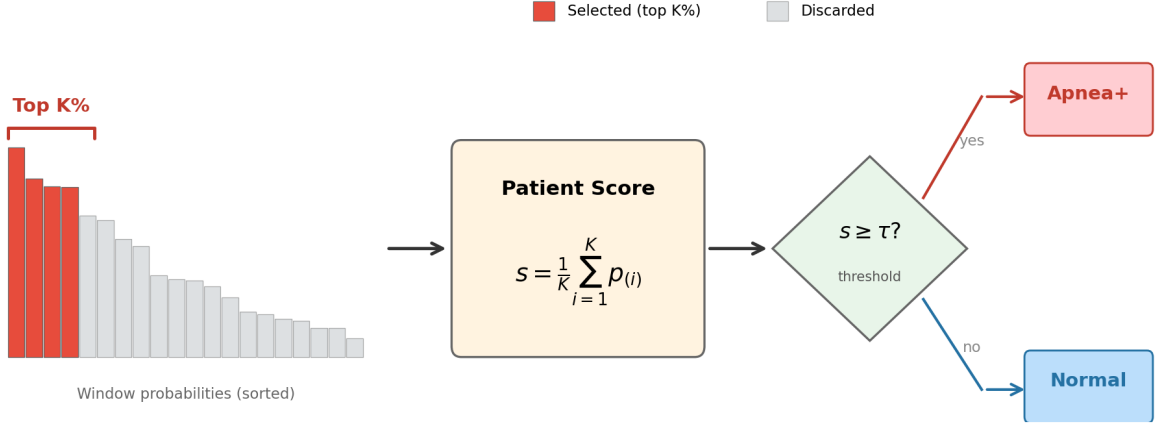


Fig. 14. Top-K aggregation converts window probabilities into a patient score. The windows are ranked, the highest K% (red bars) are selected, and their mean is computed. This reduces dilution by normal-breathing windows and enriches the aggregate for event-related evidence; it is an empirical proxy, not a direct AHI count.

The CoTELMo architecture extends the CNN baseline by incorporating temporal modeling through bidirectional LSTM layers with skip connections. The architecture draws inspiration from the TELMo (Temporal ELMo) design proposed by Alves et al. [37] for wind direction nowcasting, adapting and extending its core principles to the different domain of biomedical signal classification. The adaptation involved redesigning the convolutional front-end for 1D physiological signals, introducing a multi-resolution skip connection scheme tailored to the time scales of respiratory events in PPG, and developing the complete training and aggregation pipeline for patient-level sleep apnea screening. The resulting architecture, CoTELMo, consists of three main components: a convolutional feature extractor, two TELMo blocks, and a classification head with skip connections. Figure 15 illustrates the complete architecture.

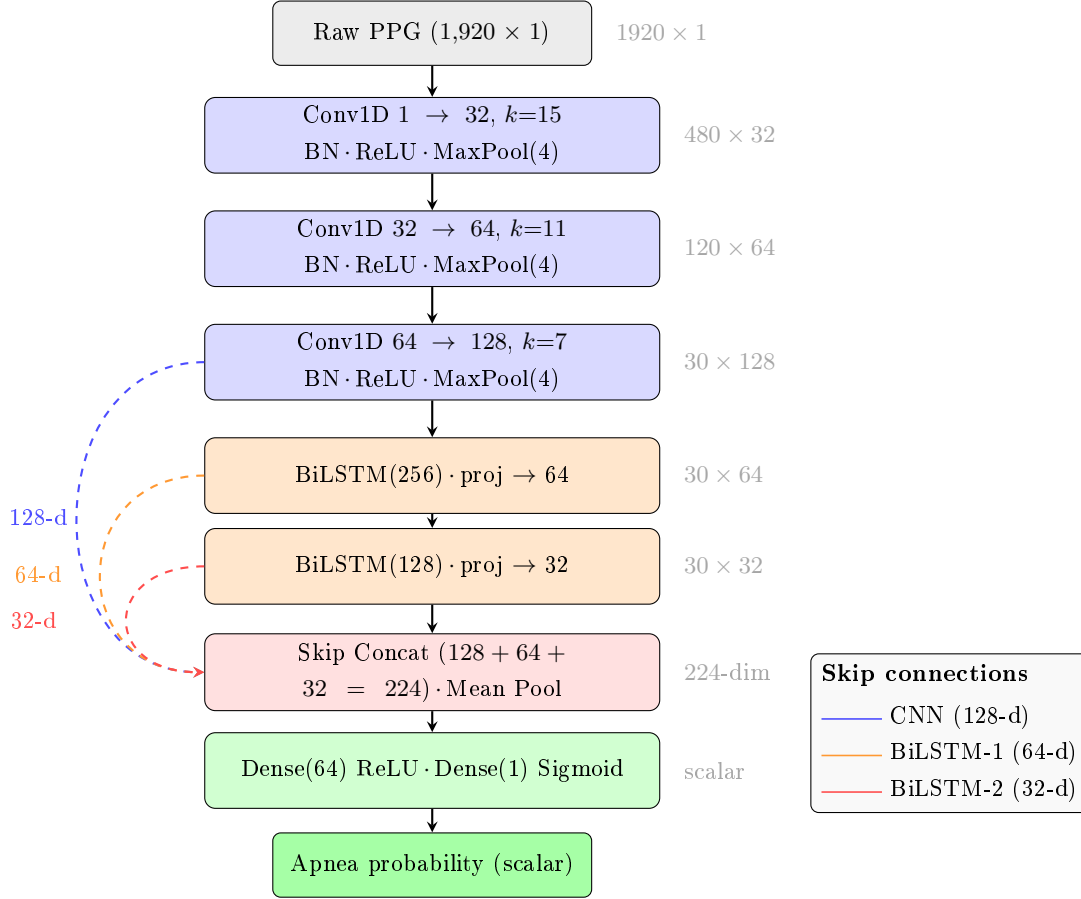


Fig. 15. CoTELMo architecture. Three convolutional blocks progressively compress the raw 1,920-sample input into a 30×128 feature sequence. Two bidirectional LSTM blocks (TELMo blocks) capture temporal dependencies at increasing context scales. Skip connections (dashed lines) concatenate the outputs of all three stages ($128 + 64 + 32 = 224$ dimensions) after mean pooling, providing the classification head with multi-resolution features spanning local pulse morphology to full-window breathing patterns. Total: ~ 1.1 M parameters.

The convolutional component consists of three blocks that progressively compress the raw 1,920-sample input (30 seconds at 64 Hz) into a sequence of 30 timesteps with 128 features. The number of blocks (three) was chosen to achieve a total temporal compression factor of $4^3 = 64$, which maps the 1,920 input samples to exactly 30 output timesteps, one per second of the 30-second window. This one-per-second resolution provides a natural time scale for the BiLSTM layers, as each timestep aggregates approximately one cardiac cycle (at a typical heart rate of 60 bpm), allowing the recurrent layers to model breathing patterns across consecutive heartbeats.

The first block uses 32 filters with kernel size 15 (corresponding to approximately 234 ms at 64 Hz, slightly longer than a typical cardiac systole), followed by batch normalization, ReLU

activation, max pooling with factor 4, and dropout. The second block uses 64 filters with kernel size 11, and the third block uses 128 filters with kernel size 7, each followed by the same normalization, activation, pooling, and dropout sequence. The decreasing kernel sizes across blocks follow the principle that early layers should capture longer-range patterns in the raw signal (spanning multiple cardiac phases), while deeper layers operate on already-compressed feature sequences where shorter kernels suffice to capture local feature interactions. The increasing filter counts (32, 64, 128) provide progressively richer feature representations as the temporal resolution decreases.

The first TELMo block takes the 30×128 CNN output and processes it through a bidirectional LSTM with 256 units per direction (512 total). The LSTM output passes through a fully connected layer that projects from 512 to 64 dimensions. The second TELMo block takes this 30×64 output and processes it through a bidirectional LSTM with 128 units per direction (256 total), followed by a projection from 256 to 32 dimensions.

The key architectural feature is the use of skip connections that concatenate representations from all three stages: the CNN output (128-dim), the first TELMo block output (64-dim), and the second TELMo block output (32-dim). After mean pooling across the 30 timesteps, these are concatenated into a single 224-dimensional feature vector. This vector passes through a classification head consisting of a fully connected layer ($224 \rightarrow 64$ with ReLU and dropout 0.5) and a final linear output layer ($64 \rightarrow 1$). The total number of trainable parameters is approximately 1,125,000.

The skip connections serve a specific purpose: they allow the classification head to access both low-level features from the CNN (local pulse morphology) and high-level temporal patterns from the BiLSTM layers (breathing regularity across the 30-second window). This multi-resolution representation is important because apnea events manifest at different time scales in the PPG signal.

The CoTELMo model was trained on the MESA dataset using 5-fold stratified cross-validation at the patient level, ensuring that all windows from a given patient appear in either the training set or the test set, but never both. Each fold used approximately 1,400 patients for training, 246 for validation (early stopping), and 411 for testing. The total number of labeled windows was approximately 2,452,790.

The loss function was Binary Cross-Entropy with logits (BCEWithLogitsLoss), with a positive class weight (`pos_weight`) computed from the training set class balance. The class weighting was necessary because, although the overall class distribution was approximately balanced (54.9% positive), individual folds could exhibit slight imbalances that would bias the model toward the majority class.

The optimizer was Adam [96] with an initial learning rate of 5×10^{-4} and weight decay of 1×10^{-4} . The learning rate was selected through preliminary experiments on a single fold: values above 10^{-3} caused training instability with loss oscillations, while values below 10^{-4} converged too slowly for the 50-epoch training budget. Weight decay of 10^{-4} was chosen to regularize the 1.1M parameters without overly constraining the convolutional filters, which need sufficient capacity to learn diverse PPG morphological patterns across the 2,056 patients in MESA.

The batch size was 512 windows, a value determined by the GPU memory constraint (12 GB). Larger batches (1,024 or 2,048) would improve gradient estimation stability but exceeded the available memory, while smaller batches (128 or 256) would increase training time without a clear benefit given the already large dataset. Gradient clipping at a maximum norm of 1.0 was applied to stabilize training, preventing the exploding gradient problem that can occur in BiLSTM architectures when processing long sequences. The learning rate scheduler was ReduceLROnPlateau with patience of 3 epochs and a reduction factor of 0.5, allowing the model to make rapid initial progress and then fine-tune more carefully as it approached convergence. Training ran for a maximum of 50 epochs with early stopping based on validation AUC (AUROC, patience = 10), with the best model checkpoint saved for each fold. Monitoring AUC rather than loss for early stopping is appropriate for binary classification because it directly measures discriminative performance and is not affected by the scale of the loss function. In practice, training converged within 15–25 epochs for most folds.

The 5 fold-specific models were saved for inference, with two distinct evaluation procedures. For MESA cross-validation, each patient was predicted only by the fold-specific model for which that patient belonged to the held-out test partition; these out-of-fold scores produced the reported MESA results. For a genuinely new or external patient, all 5 models are blind to that patient, so their window-level probabilities can be averaged before patient-level aggregation. All training was

performed on an NVIDIA GeForce RTX 3060 GPU with 12 GB of memory, using PyTorch with CUDA 12.1. Figure 16 illustrates this external-patient inference procedure.

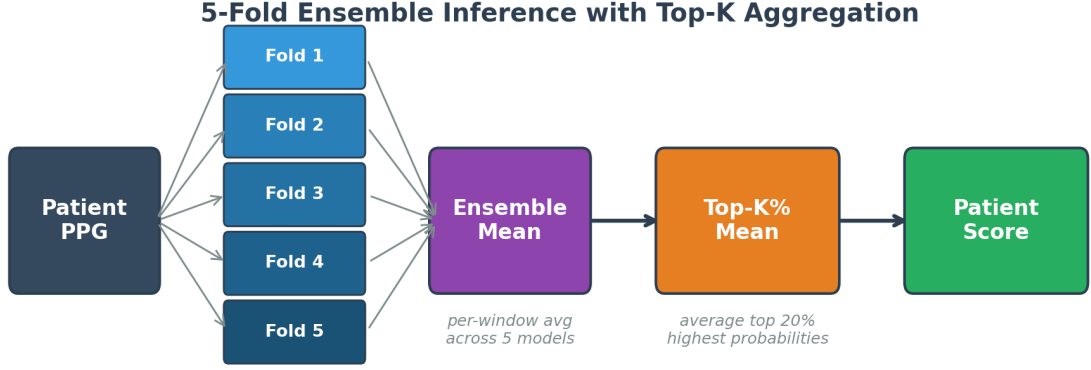


Fig. 16. Ensemble inference for a new or external patient. Because none of the 5 fold-specific CoTELMo models has seen that patient, their window probabilities are averaged before top-20% patient-level aggregation. MESA performance is instead evaluated from fold-specific out-of-fold predictions, without averaging models that trained on the evaluated patient.

Beyond binary screening, the patient-level score produced by top-K aggregation was used to derive a three-class severity label: Normal (AHI < 5), Mild (AHI 5–15), and Moderate-to-Severe (AHI ≥ 15). This mapping requires two decision thresholds, τ_1 and τ_2 , applied to the continuous score: patients scoring below τ_1 are classified as Normal, those between τ_1 and τ_2 as Mild, and those above τ_2 as Moderate-to-Severe. Both thresholds were optimized jointly within each cross-validation fold by maximizing the quadratic Cohen’s kappa on the fold’s training patients, using a grid search over the unit interval with step 0.01. Optimizing inside the fold prevents leakage: the test patients contribute neither to threshold selection nor to score calibration. Quadratic kappa was chosen as the optimization criterion because it penalizes misclassifications in proportion to their clinical distance, assigning four times the penalty to a two-step error (Moderate-to-Severe classified as Normal) compared to a one-step error (Mild classified as Normal).

4.6 Continuous AHI Estimation

While binary screening (apnea yes/no) is useful for triage, clinical practice requires a continuous estimate of severity. The AHI is the standard metric, measured in events per hour. This section describes the two-stage pipeline developed for continuous AHI estimation and the ablation study that justified the design choice.

The two-stage pipeline separates classification and regression into distinct steps. In the first stage, fold-specific CoTELMo models produce out-of-fold window probabilities for held-out MESA patients. These probabilities form a distribution that captures overall abnormality across the night. In the second stage, 12 statistical features are extracted from each patient’s probability distribution: mean, median, standard deviation, the 5th, 25th, 75th, and 95th percentiles, interquartile range, fraction above 0.5, skewness, kurtosis, and range.

These 12 features were used as input to a Support Vector Regressor (SVR) with a Radial Basis Function (RBF) kernel [97]. The SVR was evaluated using 5-fold cross-validation with out-of-fold predictions, meaning that for each fold, the CoTELMo probabilities were generated by the fold-specific model that had never seen the test patients during training. This ensures that the regression evaluation reflects the realistic scenario where both the feature extraction and regression models encounter new patients.

Feature importance analysis using Ridge regression coefficients revealed that the mean window probability was the dominant predictor (coefficient 58.1), followed by the median (20.6), the 75th percentile (12.0), the fraction above 0.5 (10.6), and the standard deviation (9.6). This ranking is clinically intuitive: AHI is a count of events, which correlates most strongly with the average probability of abnormality. Figure 17 provides an overview of the complete two-stage AHI estimation pipeline.

To justify the two-stage design, an ablation study was conducted in which the CoTELMo architecture was trained directly for AHI regression at the window level. In this direct regression approach, each window receives a regression target equal to the patient’s AHI value, and the model is trained to predict this value directly using Huber loss ($\delta = 10.0$). The patient-level prediction is obtained by aggregating window-level predictions using the top 20% mean strategy.

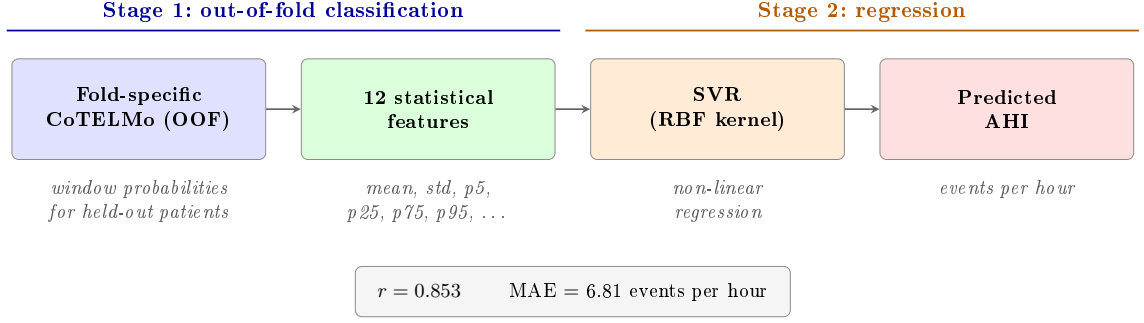


Fig. 17. Two-stage continuous AHI estimation pipeline. Stage 1 uses fold-specific out-of-fold CoTELMo probabilities, so each MESA patient is represented by a model that did not train on that patient. Stage 2 extracts twelve distribution features for an RBF-SVR that predicts AHI ($r = 0.853$, $\text{MAE} = 6.81$).

The direct regression approach achieved a Pearson correlation of $r = 0.663$ and MAE of 15.59 events per hour, worse than the two-stage pipeline ($r = 0.853$, $\text{MAE} = 6.81$). The failure of direct window-level regression stems from target ambiguity: assigning the patient’s AHI to every window creates a many-to-one mapping where windows with normal breathing receive the same high target as windows containing apnea events. This ambiguity is manageable for binary classification (the model simply learns to identify abnormal windows), but for regression it creates conflicting gradients that prevent convergence to meaningful per-window severity estimates.

4.7 Transfer Learning and Domain Adaptation

A model that only works on the dataset it was trained on has limited clinical value. This section describes the investigation of cross-dataset generalization and the strategies developed to maintain performance when the model is deployed on new sensors and populations. The transition from a model trained on controlled clinical PSG data (MESA) to a wearable consumer device (Absolute Rest) requires understanding the nature and magnitude of the domain shift, evaluating whether off-the-shelf adaptation techniques can correct it, and ultimately developing a supervised adaptation protocol that works with the minimal labeling available in clinical practice.

The MESA-trained CoTELMo ensemble was applied directly (without any adaptation) to four external NSRR datasets, after correcting PPG signal polarity for the datasets that needed it. This zero-shot evaluation measures the model’s ability to generalize without any exposure to target-domain data, and its results determine whether adaptation is necessary. Performance varied widely at $\text{AHI} \geq 15$: BestAIR achieved AUC 0.969 (above the MESA reference, but on a cohort with only

one negative subject, so the value should be read with caution), HomePAP reached 0.617, CFS reached 0.568 and HeartBEAT reached 0.499. The HeartBEAT result is limited by class composition because that dataset contains almost no normal subjects, which makes AUC estimation unreliable. The variation across datasets is itself informative: the fact that BestAIR generalizes well while CFS does not, despite both being NSRR datasets with similar clinical populations, points to sensor-level differences as the primary source of domain shift.

Analysis of the score distributions indicated that dataset-level shifts dominate the separation observed between severity groups within each individual dataset. The model is better at distinguishing which sensor produced the recording than at distinguishing severity within a recording, confirming that the domain shift operates at the level of signal characteristics rather than clinical content. This finding has a direct practical implication: calibration to a new sensor is necessary before deployment, but the underlying feature representations learned from MESA appear to be transferable.

An initial hypothesis was that differences in PPG signal polarity (inverted waveform in some sensors) could explain the shift. PPG polarity depends on whether the hardware outputs the signal as absorption (peaks up) or transmission (peaks down), which varies across manufacturers. Investigation revealed that 3 of the 4 external datasets had different polarity than MESA. However, correcting the polarity did not improve CFS performance, since CFS already had the same polarity as MESA. The domain shift therefore involves deeper differences in sensor hardware, signal conditioning, and recording protocols that cannot be resolved by simple preprocessing.

Two unsupervised adaptation methods were evaluated on CFS, neither requiring any labels from the target dataset.

Batch Normalization Adaptation (BN-Adapt) [98] recalculates the running mean and variance statistics of all batch normalization layers using target-domain data, while keeping all model weights frozen. The rationale is that domain shift may manifest as a shift in feature distributions that can be corrected by updating the normalization statistics. On CFS, BN-Adapt produced an AUC of 0.553 ($AHI \geq 15$), which is worse than the 0.568 baseline without adaptation.

CORrelation ALignment (CORAL) [99] aligns the second-order statistics (covariance matrix) of the target-domain features to match the source domain. This was applied to the 224-dimensional bottleneck features extracted by the frozen CoTELMo backbone. On CFS, CORAL achieved an AUC of 0.588, a negligible improvement of 2 percentage points over the baseline.

The failure of both unsupervised methods suggests that the domain shift operates at a level deeper than first and second-order feature statistics. The sensor-specific differences affect the learned representations in ways that cannot be corrected without supervision. Figure 18 illustrates the frozen backbone adaptation strategy that proved successful.

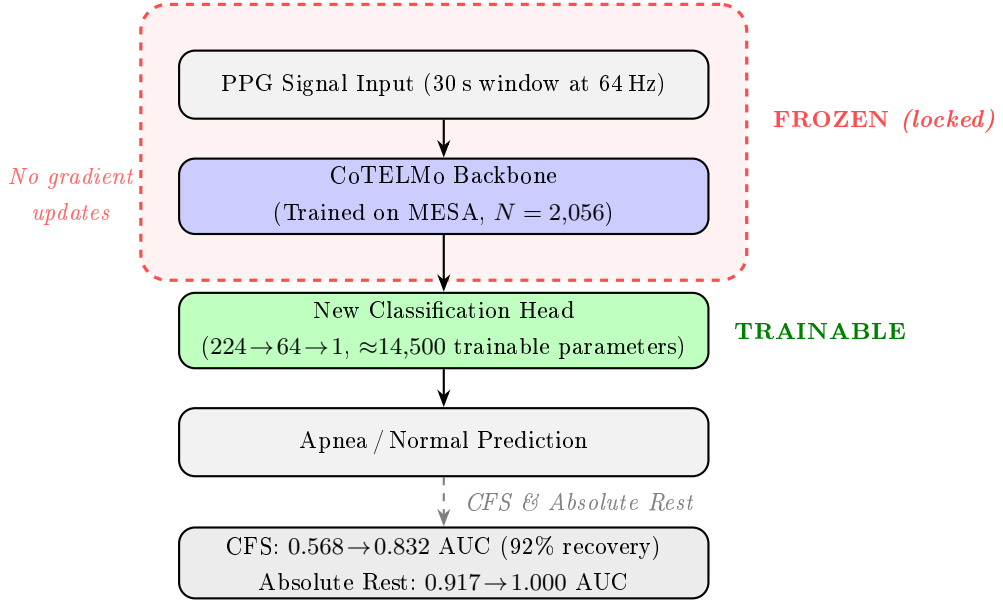


Fig. 18. Frozen backbone transfer learning strategy. The CoTELMo model trained on MESA is frozen, and only a lightweight classification head is trained on the target domain using weak patient-level labels, recovering 92% of MESA performance on CFS and reaching 100% LOO accuracy in the Absolute Rest pilot.

The supervised adaptation strategy freezes 100% of the CoTELMo model (CNN, BiLSTM layers, and all intermediate dense layers) and trains only a new classification head on the target-domain data. This approach extracts the 224-dimensional bottleneck features from the frozen backbone and passes them through a new fully connected classifier ($224 \rightarrow 64 \rightarrow 1$ with ReLU activation and dropout). Only weak patient-level labels are required, since the backbone has already learned meaningful window-level representations from the MESA strong labels. The motivation

for freezing the entire backbone, rather than fine-tuning selected layers, stems from a practical constraint: the target datasets (CFS with 324 patients, Absolute Rest with 46) are too small to reliably update the 1.1M backbone parameters without overfitting. By reducing the trainable parameter count to approximately 14,500, the risk of overfitting to the small target set is minimized while still allowing the decision boundary to shift to accommodate the new signal characteristics.

The fine-tuning protocol used Adam with a learning rate of 3×10^{-4} , weight decay of 10^{-4} , and trained for 30 epochs with early stopping. The classification head was initialized randomly for each fold, and per-fold class weights were computed from the training set class balance to compensate for the severity distribution differences between MESA and the target datasets.

On CFS (N=324), 5-fold stratified cross-validation with the frozen backbone strategy achieved an AUC of 0.832 ± 0.044 at $\text{AHI} \geq 15$, recovering 92% of the MESA training performance (0.909). At $\text{AHI} \geq 10$, the adapted model reached 0.832 ± 0.045 , and at $\text{AHI} \geq 5$ it reached 0.848 ± 0.064 . This represents an improvement of 0.264 AUC points over the unadapted baseline (0.568), using only weak labels and retraining approximately 14,500 parameters in the new classification head.

The transfer of the full pipeline (binary classification, AHI regression, multiclass grading) showed an important asymmetry. Binary screening transferred well (AUC 0.832, 92% of MESA). Multiclass severity grading was moderately preserved (quadratic kappa 0.646 vs. 0.703 on MESA). However, continuous AHI regression degraded (Pearson $r = 0.503$ vs. 0.853 on MESA), because the weak labels used for fine-tuning produce binary probability distributions that lack the graded severity information needed for accurate regression. This result supports a clear principle: screening-level classification transfers well with weak labels, but continuous severity estimation requires strong event-level labels.

4.8 Phenotype Classification

The final component of the pipeline addresses the most clinically specific task: distinguishing between OSA and CSA using PPG features alone. These two conditions have different pathophysiology and require different treatments (CPAP for OSA, ASV for CSA), making phenotype discrimination clinically necessary [17]. This task is also the most challenging from a machine learning perspective: unlike apnea detection, which benefits from a large training set (2,056 MESA

patients) and strong event-level labels, phenotype classification must work with only 46 labeled recordings from the Absolute Rest dataset, of which only 12 are pathological (6 OSA, 6 CSA). The small sample size precludes training a deep learning model from scratch and necessitates either careful feature engineering or the reuse of representations learned from the larger MESA dataset.

The initial approach to phenotype discrimination was based on the hypothesis that OSA patients would show hyperventilatory compensation, a period of elevated respiratory rate following each apnea event as the body recovers from hypoxia, while CSA patients would show absent or reduced respiratory effort. A Hyperventilation (HV) Load metric was defined as the fraction of time during which the estimated respiratory rate exceeded 30 breaths per minute, and an Event Index counted the number of discrete high-respiratory-rate episodes.

On the Absolute Rest dataset (N=40, excluding PVCs and CSA), the strict HV Load rule (30 bpm threshold) achieved 0% sensitivity for OSA detection, with all 6 Severe OSA patients classified as Normal. The HV Load values for normal subjects (0.0–5.8%) overlapped broadly with those of severe OSA patients (0.0–12.5%), providing no clinically useful separation at this threshold; a looser threshold yielded only an exploratory group-level signal with weak patient-level detection. To verify this was not a dataset-specific artifact, the same analysis was repeated on a stratified sample of 300 MESA patients (100 Normal, 100 Moderate, 100 Severe), using all 50 respiratory rate models to estimate HV Load. The results were equally negative: Normal subjects had a mean HV Load of 0.14%, Moderate 0.27%, and Severe 0.18%, with zero statistical separation between groups. Of 300 patients, only 1 was detected using the event-based approach, and 0 using the hybrid approach.

This failure to generalize demonstrated that respiratory rate-based biomarkers cannot reliably detect sleep apnea at the population level. The underlying reason is the existence of a “silent phenotype” of OSA patients who do not exhibit compensatory hyperventilation after apnea events. This finding motivated the pivot from respiratory rate features to direct classification using deep learning on the MESA dataset.

For the phenotype classification task on the Absolute Rest dataset, 26 morphological features were extracted from each recording. These features characterize the shape and regularity of the

PPG waveform across the full night, capturing patterns that differ between OSA (where respiratory effort against a closed airway causes characteristic PPG amplitude modulations) and CSA (where the absence of respiratory effort reduces PPG variability).

The features included peak amplitude statistics (mean, standard deviation, coefficient of variation), valley depth measures, inter-beat interval regularity, pulse width variability, systolic-diastolic ratio statistics, respiratory modulation power in the RIIV band, and several entropy-based complexity measures. Feature selection was performed using SelectKBest with the ANOVA F-test, optimized via Optuna [95] over 200 trials, selecting the 15 most discriminative features.

The selected features were classified using an Extremely Randomized Trees (ExtraTrees) classifier [94], chosen for its resistance to overfitting with small sample sizes and its ability to handle feature interactions without explicit engineering. The hyperparameters were optimized using Optuna: 300 estimators, maximum depth of 7, minimum samples per split of 5, and class weights of 0.530 for Normal, 3.009 for Severe OSA, and 1.916 for CSA, reflecting the 34:6:6 class imbalance.

Evaluation was performed using LOOCV on the 46 recordings (34 Normal, 6 Severe OSA, 6 CSA), where each recording was held out once while the remaining 45 were used for training. The LOOCV was repeated across 5 random seeds, and the results were averaged. The mean accuracy across the 5 seeds was $90.4\% \pm 1.7\%$, with complete separation between OSA and CSA (all 12 pathological cases correctly classified in every seed). The misclassifications occurred exclusively among Normal patients being classified as one of the pathological categories. This result was corroborated by a stratified 3-fold cross-validation that yielded $90.9\% \pm 0.9\%$.

A methodological note: the Optuna hyperparameter optimization was performed outside the LOOCV loop, which introduces a potential positive bias. To quantify this bias directly, the experiment was repeated with Optuna nested inside each LOOCV fold: for each of the 46 outer folds, 200 Optuna trials with stratified 3-fold inner cross-validation optimized the hyperparameters using only the 45 training samples, with the held-out sample never influencing the search. This fully nested procedure yielded 89.1% accuracy (41/46), compared to 90.4% for the original external-optimization protocol, a difference of 1.3 percentage points. The negligible gap confirms that external hyperparameter selection did not materially inflate the reported results.

In addition to the hand-crafted morphological features, the 224-dimensional bottleneck features from the frozen CoTELMo backbone were also evaluated for phenotype classification. This approach replaces the 26 morphological features with the learned representations, using the same LOO framework but with a new 3-class classifier head ($224 \rightarrow 128 \rightarrow 64 \rightarrow 3$ with softmax and cross-entropy loss). Training used Adam with learning rate 3×10^{-4} , weight decay 10^{-4} , and 80 epochs per LOO fold with per-fold class weights.

The CoTELMo-based approach achieved 100% LOO accuracy (46/46), with Cohen’s kappa of 1.000 and macro-averaged AUC of 1.000, surpassing the ExtraTrees classifier. All three aggregation strategies tested (mean, top 10%, top 20%) produced identical results. Analysis of the unadapted binary MESA scores already showed significant three-group separation before any fine-tuning (Kruskal-Wallis $H = 12.6$, $p = 0.0018$), with OSA scoring above Normal and CSA scoring below, suggesting that the CoTELMo backbone had learned phenotype-discriminative features during its binary training on MESA despite never being exposed to phenotype labels.

The 6 PVC recordings, held out completely from all training and optimization, were used as a robustness check. Premature ventricular contractions cause irregular PPG pulse morphology that could potentially be confused with the signal changes produced by respiratory events. Using the CoTELMo-based 3-class classifier, 5 of the 6 PVC recordings were correctly classified as Normal. The single misclassification (session 91d754e7) was predicted as CSA with a narrow margin ($P(\text{CSA}) = 0.518$ vs. $P(\text{Normal}) = 0.475$, $\Delta = 0.043$). This session had the highest window activity rate among all PVC patients ($\text{frac}_{>0.5} = 0.422$), suggesting that its severe cardiac irregularity mimicked the reduced-variability PPG pattern characteristic of central apnea.

4.9 Software and Hardware

All experiments were implemented in Python. The apnea detection pipeline (CoTELMo and all MESA-based experiments) was implemented in PyTorch. The respiratory rate regression models were implemented in TensorFlow/Keras. Classical machine learning models and feature extraction used scikit-learn [100]. Hyperparameter optimization was performed using Optuna [95] with the Tree-structured Parzen Estimator (TPE) sampler. Signal processing used SciPy and NumPy. PSG data from the NSRR was read using pyEDFlib and MNE-Python. All deep learning training was performed on an NVIDIA GeForce RTX 3060 GPU with 12 GB of VRAM, using CUDA 12.1.

4.10 Key Remarks

This chapter described the full methodological pipeline, from data acquisition through preprocessing, feature engineering, model development, and clinical validation. Five primary datasets spanning different populations, sensors, and recording conditions were used, with MESA (N=2,056) providing the foundation for deep learning through event-level respiratory annotations. Two preprocessing pipelines were developed: one at 125 Hz for respiratory rate estimation and one at 64 Hz for apnea detection, both using 30-second windows. For respiratory rate estimation, an initial attempt at direct waveform reconstruction using a U-Net architecture failed (test correlation -0.027), motivating the shift to rate estimation from extracted respiratory modulation signals. A multi-scale CNN-LSTM architecture with Optuna-optimized hyperparameters, a 50-model ensemble, and temporal smoothing post-processing achieved the final results on the combined BIDMC-CapnoBase super-dataset. The optimized classical machine learning baseline reached AUC 0.796 with logistic regression using the PPG feature candidate set and nested feature selection. The hybrid CoTELMo CNN-BiLSTM architecture (~ 1.1 million parameters) was designed to capture both local pulse morphology and temporal breathing patterns through skip connections across multiple representation levels. A two-stage regression pipeline estimated continuous AHI, motivated by the failure of direct regression under weak label noise. Among the adaptation strategies evaluated, transfer learning with a frozen backbone was the only effective approach, while unsupervised methods failed to correct the domain shift. Phenotype classification used both hand-crafted morphological features (ExtraTrees, 90.4% accuracy) and learned CoTELMo features (100% LOO accuracy), with the failure of respiratory rate biomarkers to generalize (0% sensitivity on MESA) motivating the deep learning approach.

5 Results and Discussion

Each component of the pipeline was evaluated in isolation before being assessed as a complete system. Results are reported in the order of development, from the classical baseline that defines the performance floor, through the deep learning models and their ablations, to external validation on CFS and clinical deployment on the Absolute Rest ring.

5.1 Classical Machine Learning Results

The optimized PPG-only logistic regression baseline used the 430-feature candidate set with nested feature selection, retaining 183 features in the best run. It achieved an AUC of 0.796 at the $\text{AHI} \geq 15$ threshold using nested 5-fold cross-validation. At the less stringent $\text{AHI} \geq 5$ threshold, performance dropped to 0.742, reflecting the greater difficulty of separating mild cases from normal breathing. Adding 301 advanced features (perfusion index, ODI-surrogate, nonlinear pulse rate variability, and cardiopulmonary coupling estimates) produced only a marginal improvement to 0.807, suggesting diminishing returns from further feature engineering within the classical framework.

Two class balancing strategies, SMOTE and ADASYN, were evaluated but produced no improvement over the baseline with inverse class weighting. This outcome is consistent with the literature on imbalanced clinical datasets where synthetic oversampling often introduces noise in high-dimensional feature spaces.

To contextualize the PPG-only performance, the same logistic regression framework was applied to SpO_2 -derived features (ODI3, ODI4, T90, T85, mean and minimum SpO_2 , and desaturation event statistics) extracted from the MESA oximetry channel. A combined PPG+ SpO_2 model was also evaluated. Table 6 presents the head-to-head comparison.

Table 6. Classical ML comparison: PPG-only vs. SpO_2 -only vs. PPG+ SpO_2 on MESA under the same 5-fold balanced logistic-regression protocol with mutual-information feature selection.

Task	PPG-only	SpO_2 -only	PPG + SpO_2
Binary $\text{AHI} \geq 5$ (AUC)	0.713	0.927	0.927
Binary $\text{AHI} \geq 15$ (AUC)	0.765	0.953	0.948
Multiclass macro (AUC)	0.703	0.901	0.897

Two findings emerge from this comparison. First, the SpO₂-only classifier outperforms PPG-only across all tasks, with gaps of 0.214 at AHI ≥ 5 (0.927 vs. 0.713), 0.188 at AHI ≥ 15 (0.953 vs. 0.765), and 0.198 in multiclass macro AUC (0.901 vs. 0.703). This is expected because SpO₂ directly measures oxygen desaturation, the primary physiological consequence of apnea events, while PPG captures only indirect morphological and autonomic modulations. Second, combining PPG and SpO₂ features provides no improvement over SpO₂ alone (0.948 vs. 0.953 at AHI ≥ 15). This indicates that the classical PPG features carry no complementary information beyond what SpO₂ already captures.

The clinical implications of this gap are worth noting. An AUC of 0.765 for PPG-only classification at AHI ≥ 15 means that the classifier correctly ranks a randomly chosen apnea patient above a randomly chosen normal individual in 76.5% of random pairwise comparisons. While this is better than chance, it is insufficient for clinical screening on its own, where an AUC around 0.85 or above is often considered a strong target rather than a strict minimum. The 19-point gap to SpO₂ (0.953) is large but not insurmountable: it corresponds to about 38% of the distance between random performance (0.5) and perfect classification (1.0).

This observation motivated the pursuit of learned representations through deep learning. If manual feature engineering cannot extract sufficiently apnea-discriminative information from PPG, the question becomes whether a model that operates directly on the raw waveform can learn to detect the subtle, temporally structured modulations that hand-crafted features miss. The 19-point gap to SpO₂ at AHI ≥ 15 defined the performance ceiling that the deep learning models aimed to narrow.

5.2 Deep Learning Classification on MESA

The development of the deep learning pipeline on MESA followed a progressive trajectory. Three key design decisions drove the final performance: strong event-level labelling, top-K aggregation, and the CoTELMo architecture. All MESA values are based on out-of-fold scores: each patient was predicted only by the model for whose test fold that patient was held out. Table 7 summarizes the performance at each stage.

Table 7. Progression of deep learning model development on MESA ($\text{AHI} \geq 15$).

Model	Labels	Aggregation	Window AUC	Patient AUC
CNN	Weak	Mean	—	0.621
CNN	Strong	Mean	0.771	0.823
CNN	Strong	Top-10%	0.771	0.865
CoTELMo	Strong	Top-20%	0.797	0.909
<i>Oracle ceiling</i>	—	—	—	<i>0.928</i>

The transition from weak to strong labels was the main driver of patient-level improvement. The window-level AUC rose from an unmeasured baseline to 0.771, and the patient-level AUC under mean aggregation rose from 0.621 to 0.823. This gain of 0.20 at the patient level reflects the change in supervision: weak labels assign the same patient-wide AHI category to every window, including the majority of windows that contain no apnea events, producing a noisy training signal. Strong labels mark only the windows that actually overlap with annotated events, allowing the model to learn the morphological signature of apnea rather than a patient-level prior.

Beyond the label-driven gain, the choice of aggregation strategy provided a further refinement aligned with the clinical definition of AHI. AHI is a count of discrete events per hour, so the patient signal of interest is concentrated in the minority of windows that contain apnea events. Averaging only the top 10% highest window probabilities for each patient raised the patient-level AUC from 0.823 to 0.865, an additional gain of 0.04 over mean aggregation. In a typical moderate OSA patient with an AHI of 20, apnea and hypopnea events are concentrated in a minority of the night’s windows, so the top-K window set is enriched for the diagnostic signal that mean aggregation dilutes. At 0.865 patient AUC, the CNN reaches 93% of the oracle ceiling (0.928), confirming that the window-level classifier captures most of the available discriminative content.

The CoTELMo architecture improved performance at both window and patient levels. Window-level AUC rose from 0.771 (CNN) to 0.797, while patient-level AUC improved from 0.865 to 0.909 using top-20% mean aggregation. The aggregation fraction K was re-selected for CoTELMo using the same validation procedure as for the CNN; top-20% gave the highest AUC (0.909), with top-10% nearly identical (0.908), so the result is not sensitive to the exact fraction. The optimal K varies

by task (top-10% for the CNN, top-20% for CoTELMo, and top-5% for severity grading) because it depends on how the diagnostic signal is distributed across windows for each model and label scheme; in all cases K stays small, consistent with AHI being driven by the minority of windows containing apnea events. The three innovations contributed in roughly this order of magnitude at the patient level: strong event-level labels (+0.20), top- K aggregation (+0.04), and the CoTELMo architecture (+0.04). Most of the patient-level improvement came from supervision quality; aggregation strategy and architecture provided complementary refinements that brought the model to within 2% of the oracle ceiling. The final AUC of 0.909 represents 98.0% of the oracle ceiling (0.928), indicating that the model is approaching the empirically checked maximum achievable performance given the label quality and the 30-second window granularity. The remaining 2.0% gap is likely attributable to labeling noise (inter-scorer variability in PSG annotation is itself non-trivial) and to the inherent information loss of 30-second windowing, which may split apnea events across window boundaries. Performance was consistent across all three standard AHI thresholds: 0.903 at $\text{AHI} \geq 5$, 0.901 at $\text{AHI} \geq 10$, and 0.909 at $\text{AHI} \geq 15$. Figure 19 visualizes the contribution of each innovation to the final performance.

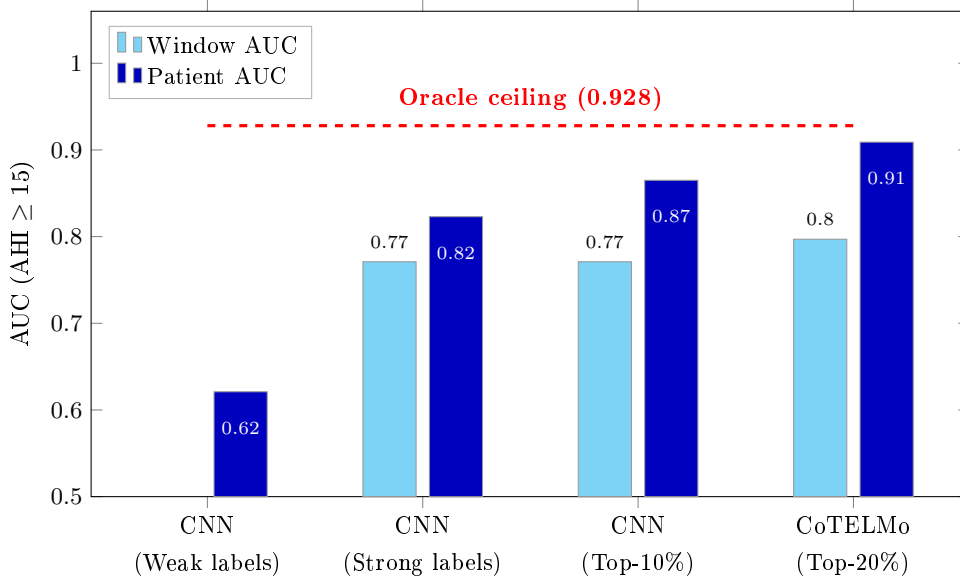


Fig. 19. Progression of deep learning model development on MESA ($\text{AHI} \geq 15$). The three innovations, strong event-level labels (+0.20), top- K aggregation (+0.04), and CoTELMo (+0.04 at fixed top-10%; selected top-20% gives 0.909), raise the patient-level AUC to 0.909, reaching 98.0% of the oracle ceiling (0.928).

While the AUC is threshold-independent, Table 8 reports sensitivity, specificity, and F1 at three representative operating points for $\text{AHI} \geq 15$. At the balanced (Youden) point, the model reached 0.796 sensitivity and 0.863 specificity. Because a screening tool should detect as many cases as possible, the threshold can be shifted toward higher sensitivity, reaching 0.90 sensitivity at 0.71 specificity, or 0.95 at 0.53. These points are read from the ROC on the evaluation scores; in deployment, the decision threshold would be calibrated on a separate validation set.

Table 8. Patient-level operating points for the CoTELMo model at $\text{AHI} \geq 15$ (top-20% mean aggregation, AUC 0.909). Thresholds are illustrative, read from the ROC on the aggregated out-of-fold scores; in deployment the decision threshold would be calibrated on a separate validation set.

Operating point	Sensitivity	Specificity	F1
Balanced (Youden)	0.796	0.863	0.820
High-sensitivity	0.901	0.709	0.816
Maximum-sensitivity	0.951	0.530	0.777

An architectural ablation was conducted to verify the necessity of the convolutional feature extraction stage. A Reshape-TELMo variant was tested, in which the raw 1,920-sample input was directly reshaped into a 30×64 sequence and fed into the TELMo blocks without any CNN front-end. This variant was designed to isolate the contribution of the convolutional blocks. The Reshape-TELMo achieved lower performance than CoTELMo, confirming that the CNN front-end provides essential feature extraction that the BiLSTM layers cannot substitute. The convolutional blocks compress and denoise the raw PPG waveform before the temporal modeling stage, and removing them forces the recurrent layers to process raw samples directly, a task for which they are not well suited given the high sampling rate (64 Hz) and the resulting long input sequences.

5.3 Multiclass Severity Grading

Before the CoTELMo architecture was developed, a TensorFlow-based three-class severity classifier (Normal / Mild / Moderate-to-Severe) was trained on MESA using weak patient-level labels. This early model achieved only 41.2% accuracy, barely above the random baseline for three classes ($\sim 33\%$). The poor performance was attributable to both the weak labeling strategy and the absence

of the temporal modeling components later introduced in CoTELMo. This result motivated the transition to PyTorch, strong event-level labels, and the hybrid CNN-BiLSTM architecture.

Beyond binary screening, the CoTELMo pipeline was evaluated for three-class severity grading (Normal with AHI < 5 , Mild with AHI 5–15, and Moderate-to-Severe with AHI ≥ 15) using dual-threshold optimization on the patient-level aggregated scores. The confusion matrix is shown in Table 9.

Table 9. Multiclass severity confusion matrix on MESA ($N = 2,056$). Rows represent true labels; columns represent predicted labels.

	Pred. Normal	Pred. Mild	Pred. Mod-Sev.
Normal ($N = 414$)	245	156	13
Mild ($N = 643$)	107	367	169
Mod-Severe ($N = 999$)	16	171	812

The overall accuracy was 69.3% with a quadratic Cohen’s kappa of 0.703, indicating substantial agreement on the Landis and Koch scale. The quadratic weighting is important because it penalizes misclassifications proportionally to their clinical distance: classifying a Severe patient as Normal (a two-step error with potentially dangerous consequences) receives four times the penalty of classifying a Mild patient as Normal (a one-step error with lower clinical impact). The top-5% aggregation strategy was selected over top-10% (which yielded a marginally higher kappa of 0.709) because it produced higher recall for the Mild category (57.1% vs. 47.6%), the clinically most ambiguous group. The kappa of 0.703 is directly comparable to the inter-scorer reliability of manual PSG scoring: the AASM programme reports 0.41 for central apnea, 0.57 for hypopnea and 0.71 for obstructive apnea [101].

The clinically most important result is the safety profile: only 1.6% of Moderate-to-Severe patients (16/999) were misclassified as Normal, meaning the model rarely produces dangerous false negatives in the highest-risk (Moderate-to-Severe) group. Moderate-to-Severe recall was 81.3%, while Normal recall was 59.2%. The asymmetry in error patterns is clinically appropriate: the model errs more frequently toward over-classification (calling a Normal patient Mild) than toward

under-classification (calling a Severe patient Normal). In a screening context, over-classification leads to a confirmatory PSG that would clarify the diagnosis, while under-classification leads to a missed diagnosis with potential health consequences.

As expected, the Mild category showed the lowest recall at 57.1%, reflecting the inherent ambiguity of this intermediate group where AHI values near the thresholds make classification difficult even for clinical scoring. A patient with an AHI of 14 and a patient with an AHI of 16 differ by only two events per hour, within the measurement uncertainty of both PSG and any automated system, yet fall into different severity categories. This boundary effect is a limitation of the categorical AHI framework rather than a limitation of the model, and it motivates the continuous AHI regression approach described in the following section. Figure 20 provides a visual representation of the confusion matrix.

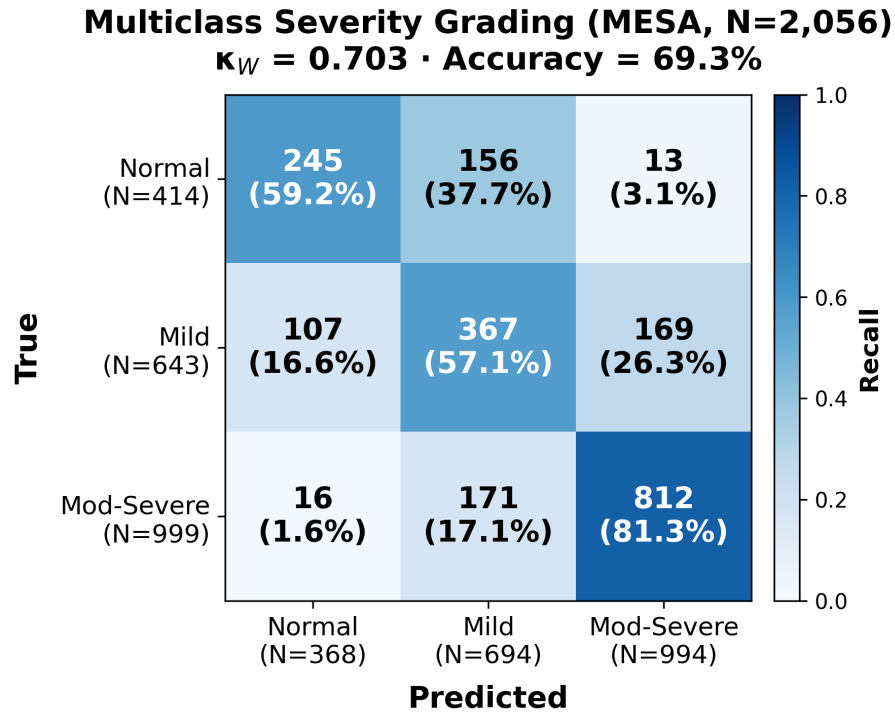


Fig. 20. Multiclass severity grading confusion matrix on MESA (N = 2,056). Color intensity represents per-class recall. The model achieves 81.3% recall for Moderate-to-Severe patients with only 1.6% dangerous false negatives (Mod-Severe classified as Normal). The Mild category shows the lowest recall (57.1%), reflecting the inherent ambiguity of this intermediate severity group.

5.4 Continuous AHI Estimation

The two-stage regression pipeline (CoTELMo window probabilities $\rightarrow$ 12 statistical features $\rightarrow$ SVR with RBF kernel) achieved a Pearson correlation of $r = 0.853$ and a Mean Absolute Error of 6.81 events per hour on MESA using 5-fold cross-validation with out-of-fold predictions. Table 10 shows the per-severity performance.

Table 10. AHI regression performance per severity group on MESA.

Severity	MAE (events per hour)	N
Normal (AHI < 5)	3.92	414
Mild (AHI 5–15)	4.26	643
Moderate (AHI 15–30)	7.11	518
Severe (AHI $\geq$ 30)	12.37	481

The higher MAE for severe patients (12.37 events per hour) is expected because AHI values in this group range from 30 to over 100, creating a wider target distribution. In relative terms, the error is proportionally smaller: an MAE of 12.4 on a mean AHI of approximately 49 represents about 25% relative error, compared to approximately 44% for mild patients (MAE 4.3 on a mean AHI near 10). This pattern of higher absolute but lower relative error in the severe group is consistent with the observation that the model captures the overall severity trend while losing precision at the extremes. For clinical use, the most important consequence is that the model rarely confuses severe patients with normal patients: a predicted AHI of 38 for a patient with a true AHI of 50 still correctly identifies the patient as severely affected.

As a preliminary baseline, the binary MESA model’s window probabilities were used directly to estimate AHI on the Absolute Rest cohort via Pearson correlation with recorded AHI values, yielding $r = 0.531$ with limited per-severity separation (mean top-10% probability ranging from 0.501 for Normal to 0.600 for Very Severe). This confirmed that a binary model alone is insufficient for continuous severity estimation and motivated the two-stage regression design.

The classical machine learning baseline used seven regression models (Ridge, Lasso, ElasticNet, SVR, Random Forest, Gradient Boosting, and k -nearest neighbours) evaluated on the 430-feature set. SVR with an RBF kernel achieved the best classical result ($r = 0.689$), which was used as the

comparison baseline throughout. Compared to this classical SVR baseline, the CoTELMo two-stage pipeline improved the correlation by 0.164 ($r = 0.853$), indicating that the learned representations capture severity-related information that hand-crafted features miss. This improvement is consistent with the design: the window-level probability distribution from CoTELMo provides a richer description of the patient’s respiratory pattern than a single vector of summary statistics. The 12 statistical features extracted from this distribution (mean, median, standard deviation, percentiles, skewness, kurtosis, and fraction above threshold) effectively compress the entire night’s information into a compact representation that is well suited for SVR regression. The direct regression ablation ($r = 0.663$, $\text{MAE} = 15.59$) supported the two-stage design, as discussed in Section 4.6. The Bland-Altman analysis revealed a mean bias of -1.40 events per hour and wide 95% limits of agreement of $[-20.9, +18.1]$. The differences also became more negative as AHI increased, indicating proportional underestimation at the severe end, consistent with the regression slope of 0.71. The fixed limits therefore summarize overall error but should be interpreted cautiously. Together, these findings support group-level triage rather than PSG replacement for individual severity grading. Figures 21 and 22 illustrate both views.

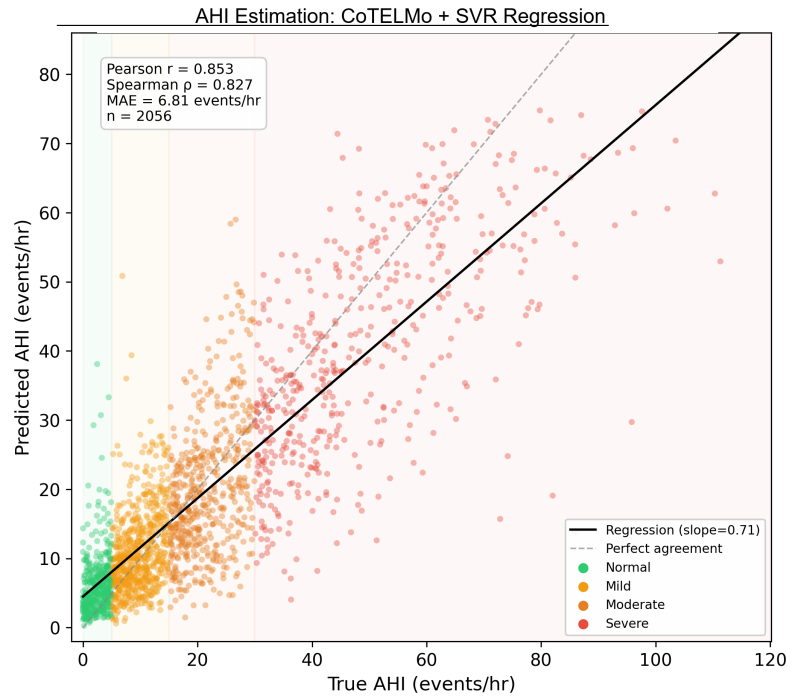


Fig. 21. Continuous AHI estimation on MESA ($N = 2,056$; $r = 0.853$, $\rho = 0.827$, $\text{MAE} = 6.81$). The regression slope of 0.71, below the perfect-agreement slope of 1, shows compression toward the mean and systematic underestimation at high AHI values.

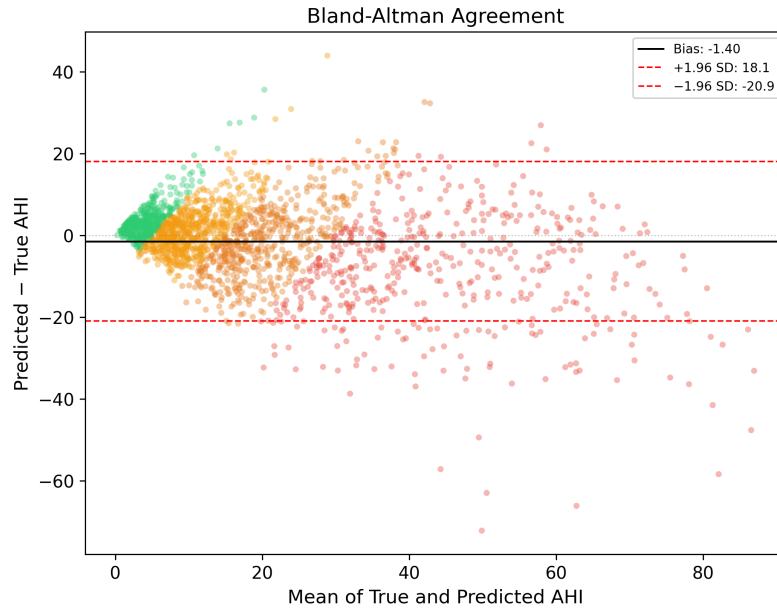


Fig. 22. Bland-Altman analysis on MESA ($N = 2,056$), with mean bias -1.40 events per hour and wide 95% limits of agreement $[-20.9, +18.1]$. Increasingly negative differences at higher mean AHI indicate proportional underestimation, so the fixed limits are descriptive rather than evidence of uniform agreement across severity.

An auxiliary Ridge model was fitted to the same standardized 12-feature representation to obtain an interpretable linear coefficient ranking, shown in Figure 23. This is not an intrinsic importance measure for the nonlinear RBF-SVR. The mean window probability dominates the Ridge analysis, indicating that average predicted abnormality carries most of its linear severity signal.

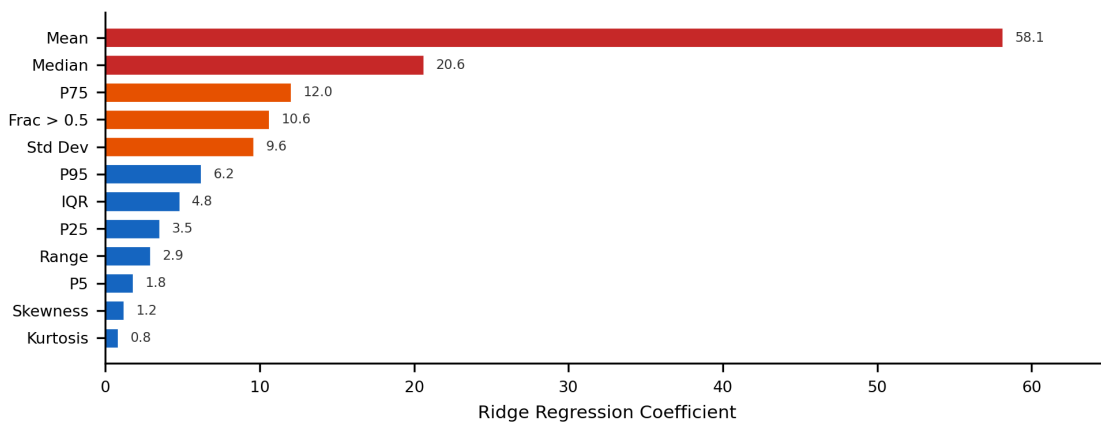


Fig. 23. Absolute standardized coefficients from an auxiliary Ridge model fitted to the 12 AHI-pipeline features. These are not RBF-SVR coefficients. Mean probability dominates by an order of magnitude (58.1), followed by median (20.6) and the 75th percentile (12.0).

5.5 Respiratory Rate Regression

The multi-scale CNN-LSTM model, trained on the combined BIDMC and CapnoBase development set with a 50-model ensemble (10 seeds $\times$ 5 folds), achieved an MAE of 1.70 bpm on BIDMC and 1.80 bpm on CapnoBase using Log-Cosh loss. These results were obtained using only PPG-derived features (RIIV, RIFV, RIPv, and three spectral features), without any ECG dependency. Both values were computed on fixed patient-level test partitions excluded from every training fold: 8 BIDMC subjects (248 windows) and 13 CapnoBase subjects (111 windows). For each test window, the 50 model outputs were averaged, followed by temporal smoothing (moving average, window size 3) where reported.

Three loss functions were compared using the same architecture and training configuration. Table 11 summarizes the results.

Table 11. Loss function comparison on the fixed held-out test partitions. BIDMC values are reported without smoothing; CapnoBase values include temporal smoothing (window size 3). All values in bpm.

Loss	CapnoBase MAE	R^2	BIDMC MAE	R^2
MSE	1.78	0.713	1.92	0.144
Huber ($\delta = 1.0$)	1.81	0.721	1.70	0.316
Log-Cosh	1.80	0.718	1.70	0.295
Grand ensemble (all 3)	1.80	0.718	1.74	0.283

MSE produced the lowest CapnoBase MAE (1.78 bpm) but also the highest BIDMC MAE (1.92 bpm), suggesting overfitting to the CapnoBase distribution. Log-Cosh provided the best balance across both datasets, matching Huber on BIDMC (1.70 bpm) while achieving a lower CapnoBase MAE (1.80 vs. 1.81 bpm). The grand ensemble combining all three loss variants offered no improvement over the single Log-Cosh model, confirming that Log-Cosh was Pareto optimal. Log-Cosh was therefore selected as the final loss function.

For comparison, the earlier two-feature model (RIIV and EDR) achieved a lower BIDMC MAE (1.37 bpm), but requires an ECG channel unavailable on wrist or ring-based wearables.

The 0.33 bpm cost of removing this dependency (1.37 to 1.70) was considered acceptable for deployment.

The final three temporal features (RIIV, RIFV, RIPV) were chosen by a feature ablation on the multi-scale architecture. Removing RIIV produced the largest MAE increase, while RIFV and RIPV gave smaller secondary contributions; among the spectral features, the dominant frequency of RIIV contributed the most. This confirms that the amplitude modulation captured by RIIV carries the strongest respiratory information, consistent with respiratory-induced intensity variation.

Tables 12 and 13 summarize the development milestones and final comparison.

Table 12. Respiratory rate model milestones (CapnoBase MAE on 13 held-out subjects).

Step	Description	CapnoBase MAE
1	Zero-shot BIDMC model (RIIV+EDR, 2ch)	5.89
2	Super-dataset, weight $5\times$ (RIIV+EDR, 2ch)	3.07
3	4-channel features (RIIV+EDR+RIAV+RIFV)	2.74
4	Multi-scale parallel convolutions (5/15/30)	1.94
5	Spectral branch (+3 scalar features)	1.87
6	Clean 6 features (RIIV+RIFV+RIPV, PPG-only) + Log-Cosh	1.80

Table 13. Respiratory rate comparison on BIDMC and CapnoBase (MAE, breaths per minute).

Method	Signal	BIDMC	CapnoBase
Pimentel et al. [46]	PPG	2.70	1.90
Bian et al. [50]	PPG	2.5 ^a	—
RRWaveNet [52]	PPG	1.66	1.59
Zhao et al. [19]	PPG + ECG	1.33	0.96
Kumar et al. [15]	PPG + ECG	— ^b	— ^b
This work	PPG only	1.70	1.80

^aCombined MAE across both datasets; per-dataset breakdown not reported [50]. ^bKumar et al. report a best 0.24 bpm (PPG+ECG); per-dataset values not reported.

These results fall within the range of recent PPG-only methods, although a direct ranking is limited by differences in datasets, windowing, and evaluation protocols. RRWaveNet [52] reports

lower MAE (1.66 vs. 1.70 on BIDMC, 1.59 vs. 1.80 on CapnoBase), but uses a larger 32-second window and a more complex multi-scale residual architecture with dedicated dilated convolution branches for each temporal scale. The Transformer-based model of Zhao et al. [19] reports still lower errors but requires an ECG channel, which limits its applicability to devices without ECG sensors. The gap between our PPG-only results and this PPG+ECG model (1.70 vs. 1.33 on BIDMC, 1.80 vs. 0.96 on CapnoBase) reflects, in part, the information cost of removing the ECG dependency. This cost is acceptable for wearable deployment, where ECG sensors are unavailable on most consumer devices.

The primary value of these results is not benchmark leadership on clinical datasets, but rather the confirmation that reliable respiratory rate can be extracted from PPG alone, a necessary prerequisite for the subsequent investigation of respiratory-rate-based biomarkers for apnea detection (which, as discussed in Section 5.8, ultimately failed).

A stratified analysis of the CapnoBase test set by respiratory rate range revealed that model performance varies with breathing frequency. In the normal range (15–20 bpm), the model achieved an MAE of 0.57 bpm (3.6% relative error). Performance degraded for slow breathing rates (10–15 bpm, MAE 1.82 bpm) and very slow rates (below 10 bpm, MAE 1.98 bpm, 20.9% relative error). The worst predictions occurred at extreme bradypnea (below 3 bpm), where individual errors exceeded 9 bpm. This pattern is explained by the training distribution: only 7.7% of the BIDMC training windows have respiratory rates below 10 bpm, compared to 43.4% in the 15–20 bpm range. The CapnoBase dataset contains a higher proportion of low-rate samples due to its surgical population (sedated patients with suppressed respiratory drive), creating a distribution mismatch that the model cannot fully compensate for despite the super-dataset training strategy.

5.6 External Validation and Domain Adaptation

The MESA-trained CoTELMo ensemble was applied directly to four external NSRR datasets without any adaptation. Table 14 summarizes the results.

The performance varied widely across datasets, from above the MESA reference on BestAIR (0.969, with one negative subject in the cohort) to near random on CFS (0.568) and HeartBEAT (0.499). The HeartBEAT result should be interpreted with caution because the dataset contains

almost exclusively moderate-to-severe patients with very few normals, making the AUC metric unreliable. The pooled AUC of 0.641 across all 760 external patients indicates a marked aggregate degradation: apparent dataset-specific performance does not translate into stable calibration across sensors, because the per-dataset score distributions are shifted relative to each other.

Table 14. Direct inference of MESA-trained CoTELMo on external datasets ($\text{AHI} \geq 15$, no adaptation).

Dataset	N	AUC	Note
MESA (internal)	2,056	0.909	Reference
BestAIR	129	0.969	Above MESA reference (cohort with 1 negative)
HomePAP	56	0.617	Moderate degradation
CFS	324	0.568	Near random
HeartBEAT	251	0.499	Unreliable (few normals)
<i>Pooled (external)</i>	760	0.641	Marked aggregate degradation

Analysis of the CoTELMo score distributions indicated that between-dataset shifts dominate the separation observed between severity groups within each individual dataset. In practical terms, the model is better at distinguishing which sensor recorded the signal than at distinguishing severity within a recording. This confirmed that the domain shift operates at the level of hardware-specific signal characteristics (sensor sensitivity, filtering, sampling precision) rather than clinical content.

Correcting PPG signal polarity (inverted waveforms in some sensors) did not resolve the shift for CFS, which already shared the same polarity as MESA. The domain shift therefore involves deeper sensor-specific characteristics that cannot be corrected by simple preprocessing transformations.

Table 15 and Figure 24 compare the three adaptation strategies evaluated on CFS ($N=324$).

Table 15. Domain adaptation results on CFS ($\text{AHI} \geq 15$, 5-fold stratified CV).

Strategy	AUC	Labels Required
No adaptation (baseline)	0.568	None
BN-Adapt	0.553	None
CORAL	0.588	None
Frozen backbone + new head	0.832 ± 0.044	Weak (patient-level)
<i>MESA (reference)</i>	<i>0.909</i>	Strong (event-level)

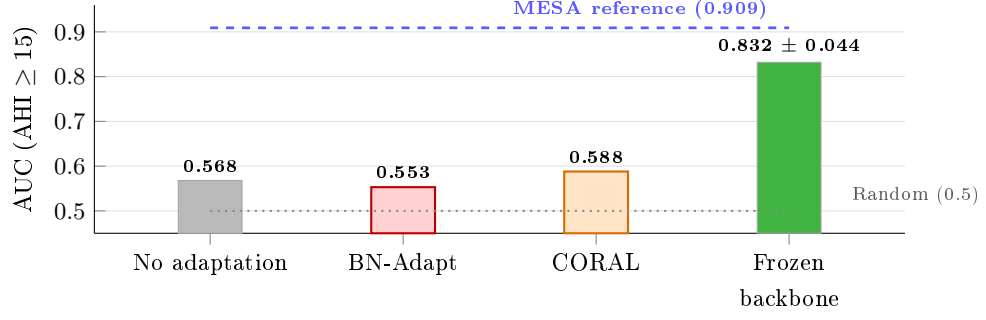


Fig. 24. Domain adaptation strategies on CFS (AHI ≥ 15). BN-Adapt degrades performance and CORAL yields only a marginal gain (+0.020); neither materially closes the domain gap. Frozen-backbone adaptation recovers 92% of MESA performance (reference AUC 0.909).

Neither unsupervised method materially closed the gap: BN-Adapt (0.553) performed worse than no adaptation, while CORAL (0.588) provided only a marginal improvement. BN-Adapt replaces source batch-normalization statistics with target-domain estimates, assuming that the shift is captured by activation means and variances. Its degradation suggests that the representations are not merely shifted or scaled. CORAL aligns feature covariance matrices and produced a small gain, but its limited effect indicates that first- and second-order alignment alone is insufficient for the MESA–CFS shift.

The supervised frozen backbone strategy was the only successful approach, achieving an AUC of 0.832 ± 0.044 at AHI ≥ 15 , which recovers 92% of the MESA training performance. This required only weak patient-level labels (binary AHI thresholding, available from any clinical record) and retraining of approximately 14,500 parameters in the new classification head, a low computational cost. The success of this strategy has a clear interpretation: the convolutional and recurrent layers of the CoTELMo backbone have learned generalizable representations of PPG temporal patterns that are relevant for apnea detection regardless of the recording device. The domain shift primarily affects the mapping from these representations to clinical decisions (the classification boundary), not the representations themselves. This is analogous to how a radiologist trained on images from one scanner can adapt to a different scanner by recalibrating their decision criteria without forgetting how to interpret the underlying anatomical structures. Performance was consistent across thresholds: 0.848 ± 0.064 at AHI ≥ 5 , 0.832 ± 0.045 at AHI ≥ 10 , and 0.832 ± 0.044 at AHI ≥ 15 . Figure 25 summarizes both the external validation results and the adaptation strategy comparison.

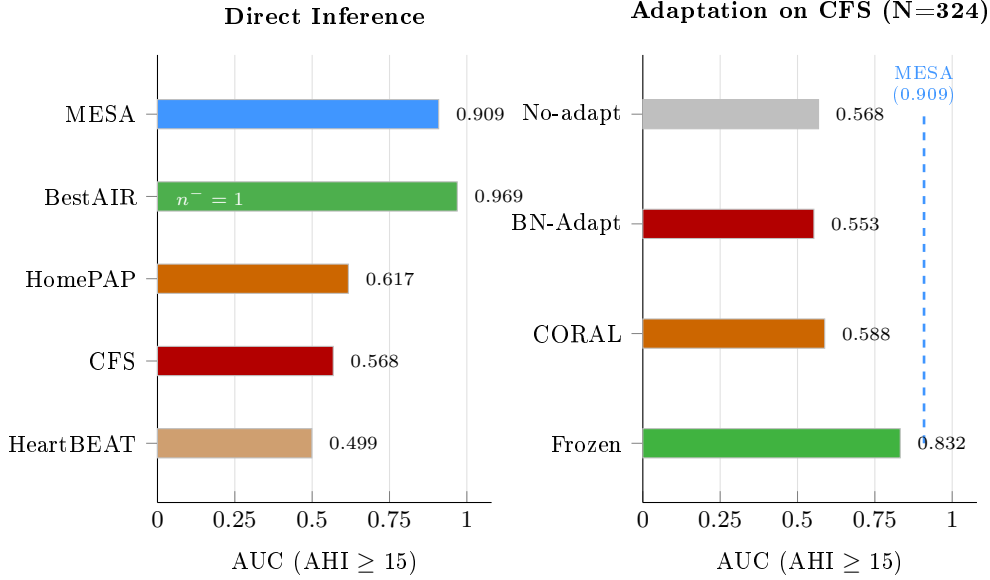


Fig. 25. External validation and domain adaptation. Left: direct inference on four external datasets; BestAIR’s AUC 0.969 is highly uncertain because the cohort contains only one negative subject, and HeartBEAT is also class-limited. Right: on CFS (N = 324), BN-Adapt degrades performance, CORAL gives a marginal gain, and supervised frozen-backbone adaptation recovers 92% of MESA performance.

An important asymmetry emerged in the transfer of different pipeline components. Binary screening transferred well (AUC 0.832, 92% of MESA). Multiclass severity grading was moderately preserved (quadratic kappa 0.646 vs. 0.703 on MESA). However, continuous AHI regression degraded ($r = 0.503$ vs. 0.853 on MESA). This degradation pattern follows a predictable gradient: tasks that require coarser distinctions (apnea vs. normal) transfer more easily than tasks that require finer-grained estimation (exact AHI value). Binary screening only needs the model to separate two broad groups, which is achievable even if the decision boundary shifts between domains. Multiclass grading requires separating three groups with two thresholds, which is more sensitive to distribution shifts. Continuous regression requires preserving the entire rank ordering of patients, which demands that the model’s output scale remain calibrated across domains, a condition that is violated when the sensor hardware produces different baseline score distributions.

This degradation occurs because weak labels produce binary probability distributions that lack the graded severity information needed for accurate regression. Strong event-level labels would allow the model to learn a graded probability scale (more events per hour = higher window probability), which would in turn produce more informative statistical features for the SVR regression stage. Without event-level labels, all windows from a positive patient receive the same probability target

regardless of whether the individual window contains an event, collapsing the intra-patient severity gradient that the regression pipeline relies on.

This result supports a practical principle: screening-level classification can be transferred with minimal supervision, but continuous severity estimation requires event-level annotations.

5.7 Clinical Validation on Absolute Rest

The Absolute Rest dataset represents the target deployment scenario: a commercial ring sensor used by elite athletes during overnight sleep. Of the 54 total recordings (36 Normal, 6 Severe OSA, 6 PVCs, 6 CSA), two Normal recordings were excluded due to missing data files, yielding 52 usable recordings (34 Normal, 6 Severe OSA, 6 PVCs, 6 CSA). The PVCs group was held out for robustness testing, so the main analyses used 46 recordings (34 Normal, 6 Severe OSA, 6 CSA). The validation proceeded in three phases: respiratory rate analysis, direct apnea inference using the MESA-trained model, and fine-tuning with LOOCV.

The CNN-LSTM respiratory rate model was applied to the Absolute Rest recordings using 30-second non-overlapping windows with temporal smoothing ($w=3$). Per-patient mean RR was 18.44 ± 1.76 bpm for Normal subjects and 20.60 ± 4.60 bpm for Severe OSA (independent t-test: $t(38) = -2.085$, $p = 0.044$; Cohen’s $d = 0.923$). While the difference reaches statistical significance, it is driven by a small subset of OSA patients with elevated mean RR, and the large standard deviation in the OSA group (4.60 vs 1.76 bpm) reflects heterogeneous respiratory patterns. A critical methodological consideration was the use of per-patient aggregation rather than per-window analysis: treating individual windows as independent observations inflated the apparent significance by orders of magnitude ($p = 9.04 \times 10^{-21}$ with pseudo-replication vs $p = 0.044$ with correct aggregation), underscoring the importance of accounting for within-subject correlation in wearable data analysis.

The MESA-trained CoTELMo ensemble was applied directly to the 46 Absolute Rest recordings (34 Normal, 6 Severe OSA, 6 CSA) without any adaptation. For the binary task of Normal vs. Severe OSA, the model achieved an AUC of 0.917 ($p = 0.0004$, top-20% aggregation), a result that contrasts with the near random performance observed on CFS (0.568). This difference suggests that the Absolute Rest PPG signal, despite originating from a consumer ring sensor, shares

relevant characteristics with the MESA clinical PPG for distinguishing severe obstructive events. One possible explanation is that the Absolute Rest ring uses a finger-based sensor, which typically produces higher-amplitude PPG signals with clearer pulse morphology than wrist-based sensors, due to the denser capillary bed in the fingertip. The MESA recordings also used finger-based pulse oximeters, so the sensor placement may contribute to the better cross-domain performance compared to CFS, which used a different sensor type.

However, for Normal vs. Any Apnea (including CSA), the AUC dropped to 0.534, near random. Further analysis revealed the cause: CSA patients received lower scores than Normal patients (mean top-10% score: 0.518 for CSA vs. 0.622 for Normal). The MESA model, trained overwhelmingly on obstructive events, had learned to detect the PPG signatures of respiratory effort against a closed airway, signatures that are absent in central apnea, where the fundamental problem is the absence of respiratory drive. This scoring reversal (CSA below Normal, Normal below OSA) is clinically meaningful because it demonstrates that the model has learned a representation that is sensitive to the presence of respiratory effort, not just to the absence of normal breathing. In obstructive events, the futile effort against a closed airway produces characteristic intrathoracic pressure swings that modulate the PPG waveform; the model has learned to detect these modulations. In central events, no effort is present, so the PPG morphology is quieter, and the model interprets this as normal or even sub-normal.

This pattern was confirmed by pairwise statistical tests (Mann-Whitney): Normal vs. OSA ($p = 0.0069$), Normal vs. CSA ($p = 0.0132$), and OSA vs. CSA ($p = 0.0087$), indicating that all three groups produce statistically distinct score distributions even before fine-tuning. Critically, OSA scores were above Normal while CSA scores were below, suggesting that the CoTELMo backbone had learned phenotype-discriminative features during its binary MESA training despite never being exposed to phenotype labels.

Fine-tuning was performed using LOOCV with the frozen CoTELMo backbone. For each of the 46 iterations, one recording was held out, and a new classification head was trained on the remaining 45 using weak labels. For the binary task of Normal vs. Severe OSA (N=40), the fine-tuned model achieved an AUC of 1.000, accuracy of 100%, sensitivity of 1.000, and specificity of 1.000. For Normal vs. Any Apnea (N=46), the results were identical: AUC 1.000 and accuracy

100%. After fine-tuning, CSA patients that had been scored below Normal in the direct inference phase now received scores above 0.995, confirming that the CoTELMo backbone already encodes CSA-discriminative features and that only the MESA-trained classifier head had failed to map them correctly.

These results must be interpreted with appropriate caution. With only 6 subjects per pathological class, the 95% confidence interval for perfect classification ranges from 54% to 100%. The LOO framework maximizes the effective training set size but can introduce optimistic bias with very small samples. Nevertheless, the consistency of perfect classification across all 46 iterations, combined with the statistical significance of the group separations in the unadapted phase, provides converging evidence that the pipeline captures meaningful clinical differences.

5.8 Phenotype Classification

The following phenotype analysis is exploratory, given the small sample size (N=46, of which 12 pathological), and the results should be interpreted as preliminary evidence rather than clinical validation. The phenotype classification task addressed the clinically important distinction between OSA and CSA on the Absolute Rest dataset (34 Normal, 6 Severe OSA, 6 CSA).

On Absolute Rest (N=40), HV Load/Event Index showed only an exploratory group-level signal and weak, threshold-dependent patient-level detection. On a 300-patient MESA stratified sample, the same biomarkers failed to generalize: 0% sensitivity and no separation across Normal, Moderate, and Severe groups (0.14%, 0.27%, 0.18%; Figure 26).

This limitation is attributed to the “silent phenotype” of OSA, where patients do not consistently exhibit compensatory hyperventilation after apnea events. A forensic analysis of individual Severe OSA patients on the Absolute Rest dataset confirmed this pattern: using the 30 bpm hyperventilation threshold, severe OSA patients had HV Load values (0.0–12.5%) that showed broad overlap with those of Normal subjects (0.0–5.8%), with no clinically useful separation between the two groups. The underlying assumption of the biomarker approach was physiologically reasonable: after an apnea event ends and the airway reopens, the patient should breathe faster than normal to compensate for the accumulated oxygen debt. This post-event hyperventilation would produce detectable respiratory rate spikes that could be counted and used as an indirect measure of apnea

frequency. However, the data revealed that severe OSA patients did not show a consistent sustained hyperventilation pattern above the 30 bpm threshold, regardless of their apnea burden. The reasons for this silent phenotype may include blunted chemosensitivity (a known consequence of chronic intermittent hypoxia in long-standing OSA), the masking effect of sleep stage on respiratory patterns (deep sleep suppresses the hyperventilatory response), and the resolution limitation of 30-second windows, which may be too coarse to capture brief 5–10 second hyperventilatory bursts.

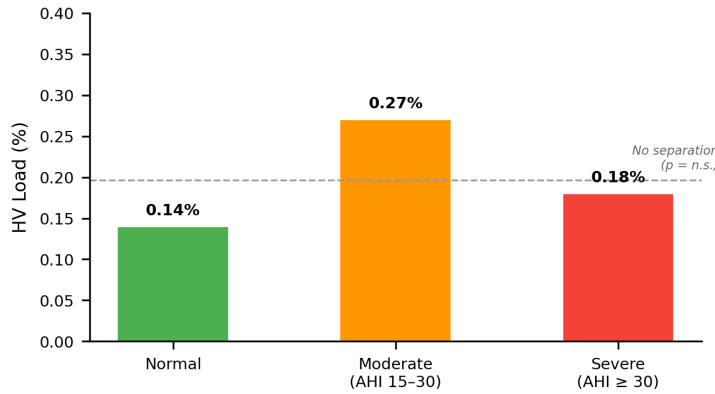


Fig. 26. HV Load (percentage of recording with hyperventilation) by AHI severity group on MESA (N = 300 stratified sample). Values show no clinically meaningful separation between groups (0.14%, 0.27%, 0.18%), confirming failure on this MESA sample. The dashed line indicates the overall mean.

While the respiratory rate regression model itself performed well on clinical benchmarks (Section 5.5), the biomarkers derived from it proved unreliable for apnea screening. This negative result is itself a contribution, as it eliminates a plausible but ultimately unproductive research direction and redirects attention toward direct morphological analysis of the PPG waveform, where the apnea-related modulations are captured implicitly by the learned CNN features rather than explicitly by respiratory rate estimation.

As an initial, lightweight phenotype baseline, the ExtraTrees classifier using 15 selected morphological features (from 26 candidates) achieved a mean LOOCV accuracy of $90.4\% \pm 1.7\%$ across 5 random seeds. A fully nested cross-validation experiment (Optuna inside each LOOCV fold; Section 4.8) yielded 89.1%, confirming that the 1.3 percentage point difference from external tuning is negligible. The confusion matrix is shown in Table 16.

Table 16. ExtraTrees phenotype classification confusion matrix (LOOCV; modal outcome across the five seeds, 41 of 46 correct, 89.1%). The $90.4\% \pm 1.7\%$ accuracy reported in the text is the mean across the five seeds. Rows: true labels; columns: predicted labels.

	Pred. Normal	Pred. OSA	Pred. CSA
Normal ($N = 34$)	29	2	3
Severe OSA ($N = 6$)	0	6	0
CSA ($N = 6$)	0	0	6

The main result is the perfect separation between OSA and CSA: all 12 pathological cases were correctly classified in every seed, with zero cross-phenotype confusion. The only misclassifications occurred among Normal patients (5 out of 34), which is clinically less concerning than missing a pathological case. A stratified 3-fold cross-validation corroborated these results ($90.9\% \pm 0.9\%$).

The potential clinical relevance of this result, if replicated on larger cohorts, relates to the direct therapeutic consequences of the OSA–CSA distinction. CPAP therapy, the standard treatment for OSA, can worsen respiratory instability in patients with predominant central apnea, who may instead require ASV or other modality-specific interventions. Currently, phenotype discrimination requires full PSG with respiratory effort bands (thoracic and abdominal), which distinguishes obstructive events (effort present, airflow absent) from central events (effort absent, airflow absent). If confirmed in a larger prospective study, the ability to infer this distinction from a single PPG channel on a wearable ring could move screening beyond binary detection toward phenotypic information that guides treatment selection before a confirmatory PSG. The present results are insufficient to establish this capability clinically, but they provide a basis for designing that larger study.

The nested 89.1% sensitivity-check estimate carries a Wilson 95% confidence interval of [77.0%, 95.3%], reflecting the uncertainty inherent in a 46-subject cohort. A separate 100-iteration bootstrap analysis confirmed the stability of the phenotype result: mean bootstrap accuracy of $92.0\% \pm 4.1\%$, 95% CI [83.6%, 99.0%]. The bootstrap OSA sensitivity was 97.2% (CI [70.6%, 100.0%]). The bootstrap Central sensitivity CI of [0%, 100%] reflects the small subgroup size ($N=6$) and should be read as a warning that per-class sensitivity estimates for CSA cannot yet be considered reliable. This does not affect the overall 90.4% result, which is driven by the correct classification of all 12 pathological cases and 29 of 34 Normal patients.

The robustness of the LOOCV protocol was assessed by comparing it against stratified 3-fold and 5-fold cross-validation across five random seeds. Stratified 3-fold yielded $90.9\% \pm 0.9\%$, consistent with the LOOCV estimate. Stratified 5-fold showed higher variance: accuracy ranged from 84.8% to 93.5% across seeds, with one seed producing OSA recall of 5/6 rather than 6/6. This higher variance under 5-fold CV, where smaller training sets are used, confirms that LOOCV is the more appropriate protocol for this sample size: it maximizes training set size for each iteration (45 recordings per fold) and produces more stable estimates across seeds.

One-vs-Rest AUC values (averaged across five seeds) were computed for each class: Normal AUC 0.948, OSA AUC 0.987, and CSA AUC 0.992. The macro-averaged AUC was 0.975. The pairwise OSA-vs-Central AUC was 1.000 across all seeds, confirming complete probabilistic separation between the two pathological phenotypes in the morphological feature space. This is a stronger result than the accuracy alone: not only are all 12 pathological cases classified correctly, but their predicted probability distributions are non-overlapping, leaving no ambiguous cases near the decision boundary.

The choice of $K=15$ features was validated by a sweep from $K=3$ to $K=26$ using the same LOOCV protocol. Accuracy at $K=3$ was 71.7%, $K=5$ was 78.3%, $K=7$ was 87.0%, $K=10$ was 87.0%, $K=15$ was 89.1%, $K=20$ was 87.0%, and $K=26$ was 89.1%. Accuracy plateaus at $K=7$ and does not increase monotonically beyond that point, and the $K=15$ selection lies within this stable plateau region. This pattern indicates that the chosen feature count is not an artifact of overfitting to the cross-validation splits: adding or removing features in the range $K=7$ to $K=26$ produces equivalent or lower accuracy, so the result is not sensitive to the precise value of K .

A learning curve analysis varied the proportion of MESA recordings used for pre-training the feature extraction pipeline from 0% to 100%. At 0% MESA (training exclusively on the 46 Absolute Rest recordings), the morphological feature classifier achieved 89.1% LOO accuracy, identical to the sensitivity-check lower bound of the full pipeline. The result remained comparable across all MESA proportions (10%, 25%, 50%, 75%, 100%), with no consistent improvement as more MESA data was added. This finding indicates that the handcrafted morphological features capture device-specific and patient-specific PPG waveform characteristics that are already distinguishable from the Absolute Rest recordings alone, without any pre-training contribution from MESA. The prac-

tical implication is that the ExtraTrees morphological feature classifier is self-contained within the Absolute Rest cohort, while the CoTELMo feature-based classifier (100% LOO accuracy) genuinely benefits from the MESA-learned deep representations. The two approaches are therefore complementary: morphological features provide a lightweight, dataset-independent baseline, whereas the CoTELMo backbone provides richer representations that improve discrimination at the cost of requiring a large pre-training corpus.

SMOTE oversampling was also evaluated for the phenotype classification task, with threshold tuning across the range 0.30–0.60. The best SMOTE result (87.0% at threshold=0.30) was lower than the ExtraTrees result without SMOTE (90.4%), consistent with the MESA binary classification finding that synthetic oversampling provides no benefit in high-dimensional feature spaces with small clinical samples (Section 5.1). Inverse class weighting was retained as the class-imbalance correction strategy.

The handcrafted-feature approach was also tested on a 4-class variant that included PVCs as a fourth category alongside Normal, OSA, and CSA. The 4-class LOO accuracy was 80.8%, with per-class recall of OSA 100% (6/6), CSA 83.3% (5/6), PVCs 50.0% (3/6), and Normal 28/34. A nested sensitivity check produced 78.8% accuracy, with Normal recall of 27/34. The PVC recall of 50% is substantially lower than the CoTELMo-based 3-class model’s effective PVC handling (5/6 correctly classified as Normal in the robustness test). This comparison highlights a limitation of the morphological feature approach: the handcrafted features do not reliably distinguish PVC-induced PPG irregularities from pathological apnea patterns, whereas the CoTELMo backbone’s learned representations appear more discriminative for PVC morphology.

The 224-dimensional CoTELMo bottleneck features, extracted from the frozen backbone and classified with a new 3-class head, achieved 100% LOO accuracy (46/46), Cohen’s kappa of 1.000, and macro-averaged AUC of 1.000. The confusion matrix showed an error-free diagonal. Per-class confidence analysis revealed that even the least-confident predictions maintained clear class separation: the lowest Normal probability for a Normal patient was 0.485 (still the dominant class), the lowest OSA probability for an OSA patient was 0.746, and the lowest CSA probability for a CSA patient was 0.859. Figure 27 compares the confusion matrices of both phenotype classification approaches.

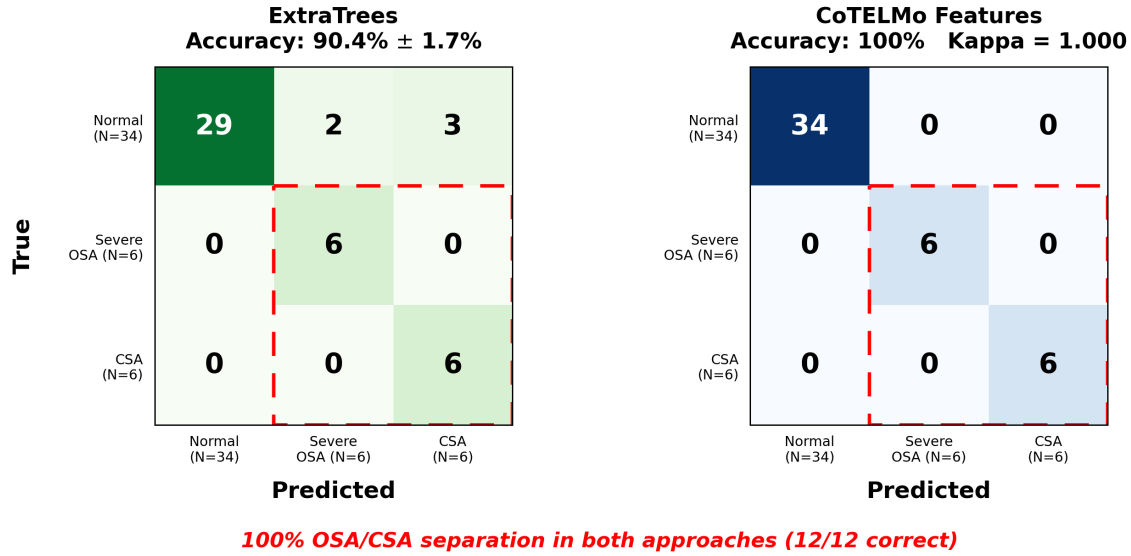


Fig. 27. Phenotype classification confusion matrices (LOOCV, $N = 46$). Left: ExtraTrees classifier on 15 morphological features; the matrix shows the modal outcome across five seeds (41/46 correct, 89.1%), while the title reports the mean accuracy across seeds ($90.4\% \pm 1.7\%$). Right: CoTELMo bottleneck features (100% accuracy). Both approaches achieve complete separation between OSA and CSA (dashed box), with all 12 pathological cases correctly classified. The only misclassifications occur among Normal patients in the ExtraTrees approach.

A complementary analysis tested the specificity of the RR model across all four Absolute Rest pathology groups (Normal, Severe OSA, CSA, PVCs). The respiratory rate variability (coefficient of variation of per-window RR) showed no significant differences across groups ($p = 0.913$, one-way ANOVA), with mean RRV scores of 22.0% (Normal), 22.9% (Severe OSA), 21.6% (CSA), and 21.6% (PVCs). This result shows no statistically significant difference in variability estimates across pathology groups, suggesting that the model's error does not depend strongly on the underlying condition, though the study is likely underpowered to confirm equivalence.

The 6 PVC recordings, held out from all training, were classified using the CoTELMo-based 3-class model. Five of six were correctly identified as Normal. The single misclassification (session 91d754e7, predicted as CSA with $P(\text{CSA}) = 0.518$ vs. $P(\text{Normal}) = 0.475$) had the most severe cardiac irregularity among PVC patients, with a pattern mimicking the reduced-variability PPG signal characteristic of central apnea. This misclassification is clinically explainable: PVCs produce irregular cardiac timing that disrupts the normal pulse-to-pulse variability of the PPG waveform. When PVCs are frequent and regular (bigeminy or trigeminy patterns), the resulting PPG mor-

phology can resemble the reduced variability seen in CSA, where the absence of respiratory effort produces a monotonous signal pattern.

This result has practical implications for deployment. PVCs are common in adults (more so in the elderly and in patients with cardiovascular disease) and are markedly more frequent in sleep apnea patients, where they have been reported in up to two-thirds of cases and the repetitive cycles of hypoxia and reoxygenation promote cardiac irritability [22]. A screening tool that consistently misclassifies PVC patients as having CSA would produce an unacceptable false-positive rate. The fact that only 1 of 6 PVC patients was misclassified, and with marginal confidence (probability difference of only 0.043), suggests that the model is largely robust to cardiac arrhythmias. However, clinical deployment should include a pre-screening step for significant arrhythmias, either through ECG or through PPG-based arrhythmia detection algorithms, to flag recordings where the apnea classification may be unreliable.

5.9 Comparison with State of the Art

Table 17 positions the results of this work within the broader landscape of automated sleep apnea detection methods.

Table 17. Comparison with published sleep apnea detection methods (moderate-to-severe apnea unless noted). Reported scores use different metrics (AUC, accuracy, F1, ICC, and κ) and are not directly comparable; the metric is annotated for each value.

Study	Signal	N	Reported performance	AHI agreement	TL	Pheno.
Gutiérrez-Tobal [65]	SpO ₂	320	0.874 (acc)	—	No	No
OxiNet [67]	SpO ₂	12,923	0.92–0.96 (ICC) ^a	0.94 (ICC)	Yes	No
Mostafa et al. [102]	SpO ₂	8+25	0.85–0.98 (acc)	—	No	No
Urtnasan et al. [60]	ECG	82	0.96 (F1)	—	No	No
Radha et al. [20] ^b	PPG	60	0.65 (κ)	—	Yes	No
Hayano et al. [68]	PPG	41	0.84 (AUC)	0.81 (r)	No	No
Jiang et al. [69]	PPG	59	0.90 (acc)	—	No	No
This work	PPG	2,056	0.909 (AUC)	0.853 (r)	Yes	Yes

^a OxiNet reports an intraclass correlation (ICC) for AHI agreement; no AUC is reported.

^b Radha et al. address sleep-stage classification (κ for four-class staging), not apnea detection; listed for cross-signal transfer-learning context only.

Several observations emerge from this comparison. First, SpO₂-based methods [65, 67, 102] report strong performance, consistent with oxygen desaturation being the strongest single biomarker

for sleep apnea; their reported scores, however, use different metrics (accuracy and ICC) from the AUC reported here, so this is a qualitative observation rather than a direct ranking. The gap between SpO₂ and PPG is expected, given that desaturation is a direct consequence of apnea events, while PPG captures only indirect autonomic and mechanical modulations. However, this gap must be weighed against the hardware trade-off: SpO₂ requires dual-wavelength sensors that are absent from most consumer wrist-worn devices, while PPG is available on virtually every smartwatch and fitness tracker. The question is not whether PPG can match SpO₂ in accuracy, but whether it can achieve sufficient accuracy for population-level screening, a task where high sensitivity and moderate specificity are acceptable because the goal is to identify candidates for confirmatory PSG rather than to replace PSG.

Second, CoTELMo achieves strong PPG-only performance (AUC 0.909) on a large public cohort (N=2,056). Earlier studies range from small cohorts, such as Hayano et al. [68] with 41 patients and Jiang et al. [69] with 59, to the 1,375-subject PPG study by Jothi et al. [70], although their targets and reported metrics differ. The scale and diversity of MESA strengthen the evaluation, while patient-level out-of-fold prediction prevents patients from contributing to the training of the model used to evaluate them. Performance was also consistent across AHI thresholds: 0.903 at AHI ≥ 5 , 0.901 at AHI ≥ 10 , and 0.909 at AHI ≥ 15 .

Third, this work integrates three capabilities in a single PPG-based pipeline: (1) cross-dataset transfer learning with quantified recovery rates, evaluated on CFS (N=324); (2) continuous AHI regression from raw PPG achieving higher correlation than Papini et al. ($r = 0.853$ vs. $r = 0.61$) [72]; and (3) a preliminary three-class subject-level phenotype analysis on single-wavelength wearable-ring PPG (N=46). The third capability extends the finger-PPG central-versus-obstructive precedent established by Massie et al. [18] to a different sensor and task, but its small sample precludes definitive conclusions. If confirmed in larger cohorts, phenotype information could help prioritize targeted PSG assessment.

The OxiNet model [67] represents the closest methodological comparison: it uses a hybrid CNN-CRNN architecture on a large multi-site dataset with cross-dataset evaluation, and reports continuous AHI estimation (ICC = 0.94 on SHHS, 0.94 on MESA). However, OxiNet uses the

SpO₂ signal rather than raw PPG, operates on the direct measure of desaturation rather than the indirect PPG modulations, and does not attempt phenotype classification.

5.10 Key Remarks

This chapter presented results across the full pipeline, from respiratory rate estimation through apnea detection to clinical deployment. The multi-scale CNN-LSTM ensemble achieved competitive respiratory rate estimation (1.70 bpm on BIDMC, 1.80 bpm on CapnoBase) using only PPG-derived features, with Log-Cosh loss providing the best cross-dataset balance and RIIV identified as the dominant feature through ablation analysis. A stratified evaluation revealed degraded performance at extreme respiratory rates below 10 bpm, attributable to training set imbalance. The matched signal comparison established a PPG-only floor of AUC 0.765 at AHI ≥ 15 under the same 5-fold cross-validation protocol used for SpO₂ and PPG+SpO₂, with a 19-point gap to SpO₂-based methods (0.953) and no benefit from PPG+SpO₂ fusion, confirming that classical PPG features carry no complementary information beyond oximetry. The CoTELMo deep learning model narrowed this gap, achieving AUC 0.909 on MESA (N=2,056), which represents 98.0% of the oracle ceiling. Three technical innovations drove this improvement, in order of magnitude: strong event-level labels (weak $\rightarrow$ strong: +0.20 patient AUC, the largest gain), top-K aggregation (mean $\rightarrow$ top-K: +0.04 patient AUC), and the hybrid CNN-BiLSTM architecture (CNN $\rightarrow$ CoTELMo: +0.04 patient AUC). The two-stage AHI regression pipeline achieved $r = 0.853$, outperforming both classical features ($r = 0.689$) and direct deep learning regression ($r = 0.663$). External validation on CFS (N=324) revealed domain shift (AUC 0.568 without adaptation), which the supervised frozen backbone protocol resolved (AUC 0.832, 92% recovery), establishing CFS as the primary external validation for cross-sensor transfer. As a preliminary pilot on the Absolute Rest ring (N=46), binary classification reached 100% LOO accuracy, and an exploratory phenotype classifier achieved 100% LOO accuracy for three-class discrimination (Normal/OSA/CSA) using CoTELMo features and 90.4% using ExtraTrees; these pilot results require validation on larger cohorts. The failure of respiratory rate biomarkers to generalize (0% sensitivity on MESA) confirmed that deep learning on raw PPG captures apnea-related information that is inaccessible to traditional signal processing approaches.

6 Conclusion

The four research questions posed in Chapter 1 provide the organizing structure for the conclusions below. Each question is assessed in turn against the experimental evidence, followed by a statement of the main contributions, a discussion of limitations, and directions for future work.

6.1 Summary and Answers to Research Questions

This work developed a pipeline for respiratory analysis and sleep apnea detection from PPG signals captured by wearable devices. The pipeline spans four levels of clinical utility: respiratory rate regression, binary apnea screening, continuous severity estimation, and phenotype classification. It was trained on large-scale polysomnographic data (MESA, $N=2,056$), validated across four independent cohorts, and deployed on a commercial wearable ring for elite athletes. The four research questions formulated in Chapter 1 are addressed below.

Regarding **RQ1**, whether deep learning models can extract reliable respiratory rates from single-lead PPG signals in the presence of real-world noise and artifacts, the answer is yes. The multi-scale CNN-LSTM model, trained as a 50-model ensemble on the combined BIDMC and CapnoBase development set, achieved an MAE of 1.70 bpm on the fixed BIDMC test set and 1.80 bpm on the fixed CapnoBase test set using only PPG-derived respiratory modulations (RIIV, RIFV, RIPV, and spectral features), with no ECG dependency. These results fall within the range of recent PPG-only methods, within 0.04–0.21 bpm of the best published methods that use larger windows or more complex architectures, although a direct ranking is limited by differences in datasets and evaluation protocols. The signal quality index filter and multi-feature fusion were sufficient for handling real-world artifacts in these datasets, supporting the conclusion that reliable respiratory rate extraction from single-lead PPG is achievable with current deep learning methods.

Regarding **RQ2**, whether a deep learning architecture trained on large-scale polysomnographic data can achieve clinically useful sleep apnea screening from PPG alone, the answer is yes, with caveats. The CoTELMo architecture achieved an AUC of 0.909 at $AHI \geq 15$ on MESA, representing 98.0% of the oracle ceiling (0.928). This constitutes strong PPG-only performance on a large public cohort ($N=2,056$), although direct ranking against prior studies is limited by differences in targets and protocols. In the matched signal comparison, the PPG-only classical baseline (AUC 0.765) established a 19-point gap to SpO_2 -based methods (AUC 0.953), and CoTELMo narrowed this gap

by 77% (from 0.188 to 0.044). However, the remaining gap to SpO₂ confirms that PPG captures only indirect apnea signatures, making it suitable for population-level screening but not yet a replacement for oximetry-based diagnostics. The continuous AHI regression ($r = 0.853$, MAE = 6.81 events per hour) further demonstrates that the pipeline provides clinically interpretable severity information beyond binary screening. Multiclass severity grading achieved a quadratic Cohen’s kappa of 0.703, comparable to PSG inter-scorer reliability.

Regarding **RQ3**, whether the learned representation transfers across different sensors and patient populations, the representation transfers partially. Direct inference on four external NSRR datasets revealed dataset-dependent performance ranging from above the MESA reference on BestAIR (AUC 0.969, on a cohort with only one negative subject) to near random on CFS (AUC 0.568) and HeartBEAT (AUC 0.499). Pooled across all 760 external patients, the AUC dropped to 0.641, indicating that per-dataset performance does not translate into stable cross-sensor calibration. Domain shift analysis confirmed that the CoTELMo embeddings capture sensor-specific characteristics more strongly than clinical severity, with between-dataset shifts dominating the separation observed between severity groups within each individual dataset. Unsupervised adaptation methods (BN-Adapt and CORAL) did not materially correct this shift. However, supervised fine-tuning with a frozen backbone recovered 92% of MESA performance on CFS (AUC 0.832 ± 0.044) using only weak patient-level labels and retraining approximately 14,500 parameters. This supports a practical adaptation protocol: binary screening transfers effectively with minimal supervision, but continuous AHI regression ($r = 0.503$ on CFS vs. 0.853 on MESA) requires event-level annotations to recover comparable performance.

Regarding **RQ4**, whether it is possible to distinguish OSA and CSA phenotypes using only PPG morphological features, the answer is yes, with the important caveat that this constitutes a preliminary exploration rather than a validated clinical capability. Both the ExtraTrees classifier with handcrafted morphological features ($90.4\% \pm 1.7\%$ LOOCV accuracy) and the CoTELMo feature-based classifier (100% LOOCV accuracy) achieved complete OSA–CSA separation with zero cross-phenotype confusion. The 6 OSA and 6 CSA patients were correctly classified in every leave-one-out iteration across all random seeds, though the small sample size ($N=46$, of which 12 pathological) precludes clinical conclusions and requires validation on larger independent cohorts

before this result can be considered reliable. Relative to prior finger-PPG central-versus-obstructive analysis [18], this pilot evaluates a distinct three-class subject-level task using a single-wavelength wearable ring. The two classification approaches are complementary in an important respect: a learning curve analysis showed that the ExtraTrees morphological feature classifier achieves equivalent accuracy (89.1%) even when trained on Absolute Rest data alone, without any MESA pre-training, indicating that the handcrafted features capture device-specific PPG waveform characteristics that are sufficient for discrimination within this cohort. The CoTELMo feature-based classifier, by contrast, relies on the rich representations learned from 2,056 MESA recordings. The finding that the MESA-trained CoTELMo backbone had encoded phenotype-discriminative features despite never being exposed to phenotype labels during training suggests that obstructive and central apnea produce different PPG morphological signatures that deep learning can discover without phenotype-level supervision.

6.2 Contributions

The main contributions of this work are summarized as follows. First, the CoTELMo architecture for PPG-based apnea screening: a hybrid CNN-BiLSTM architecture, adapted from the TELMo framework [37] and redesigned for biomedical signal classification, that achieves AUC 0.909 on MESA ($N = 2,056$). This represents strong PPG-only performance on a large public cohort. The model operates on raw 30-second PPG windows at 64 Hz and produces patient-level predictions through a clinically-motivated top-K aggregation strategy. Second, a top-K aggregation strategy that enriches the patient score for the minority of windows carrying event-related evidence. Averaging the top 10% highest window probabilities raised patient-level AUC by 0.04 over mean aggregation (from 0.823 to 0.865), refining the larger gain that came from event-level supervision (which by itself raised patient AUC from 0.621 to 0.823). Third, a two-stage AHI regression pipeline (CoTELMo window probabilities $\rightarrow$ statistical features $\rightarrow$ SVR) that achieves $r = 0.853$ and $\text{MAE} = 6.81$ events per hour, outperforming both classical feature-based regression ($r = 0.689$) and direct deep learning regression ($r = 0.663$). Fourth, a systematic characterization of PPG domain shift across multiple NSRR datasets and a frozen-backbone adaptation protocol evaluated on CFS ($N=324$), recovering 92% of source-domain performance with weak labels and approximately 14,500 trainable parameters. Fifth, a secondary pilot on the Absolute Rest wearable ring ($N=46$) extends prior finger-PPG phenotype work to three-class subject-level discrimination using a single-

wavelength sensor: direct inference achieved AUC 0.917 for severe OSA, while ExtraTrees (90.4%) and CoTELMo features (100%) achieved complete OSA–CSA separation. These pilot results require validation on larger cohorts. Sixth, robustness to PVC cardiac arrhythmias was evaluated on Absolute Rest (5/6 correctly classified as Normal).

6.3 Limitations

Despite the encouraging results, several limitations must be acknowledged, each with specific implications for clinical deployment and future research.

The main limitation is the small sample size for pathological classes in the Absolute Rest dataset. With only 6 OSA and 6 CSA patients, the 95% confidence interval for perfect classification ranges from 54% to 100% (Clopper-Pearson exact binomial). While the consistency of results across all LOO iterations and the convergence of two independent classification approaches (ExtraTrees and CoTELMo features) provide supporting evidence, the clinical reliability of phenotype classification cannot be established without larger samples. The phenotype classification results are based on a small convenience sample and should be interpreted as hypothesis-generating rather than confirmatory. A power analysis indicates that demonstrating 90% sensitivity with a 95% confidence lower bound of 70% would require at least 29 subjects per pathological class, roughly 5 times the current sample size.

The Absolute Rest validation cohort consists exclusively of elite athletes, a population that differs from the general sleep apnea demographic in terms of age, body mass index, cardiovascular fitness, and comorbidity profile. Athletes typically have lower resting heart rates, higher heart rate variability, and better peripheral perfusion than the general population. These physiological differences may make the PPG signal both stronger (better signal-to-noise ratio due to perfusion) and more atypical (resting bradycardia could be confused with apnea-related bradycardia). Whether the pipeline maintains its performance on older, less fit individuals with more complex medical histories remains an open question that requires a dedicated validation study.

Only severe OSA ($AHI \geq 30$) was represented in the Absolute Rest dataset. The pipeline’s ability to detect mild-to-moderate OSA (AHI 5–30) on wearable-grade signals has not been directly validated, although the MESA results demonstrate strong performance across all severity levels in

clinical-grade recordings. Mild OSA is both the most prevalent form and the hardest to detect, because the PPG modulations produced by mild events are smaller in amplitude and shorter in duration, potentially falling below the detection threshold of a consumer-grade sensor with lower signal quality than clinical PSG hardware.

The regression pipeline degrades when only weak patient-level labels are available ($r = 0.503$ on CFS vs. 0.853 on MESA). This is a fundamental limitation of the two-stage approach: the first stage produces binary probability distributions that lack the graded severity information needed for accurate continuous AHI estimation without event-level annotations. In practice, this means that deploying the full pipeline (including AHI regression) on a new device requires not just patient-level diagnoses but event-level polysomnographic annotations for calibration, the same annotation burden that the pipeline aims to reduce.

The model was trained and evaluated exclusively on supine or uncontrolled sleep positions. The MESA recordings include natural position changes during sleep, but specific positional analysis (supine vs. lateral vs. prone) was not performed. Positional OSA, where apnea events occur predominantly in the supine position, is a common clinical subtype that may present different PPG patterns depending on body orientation. The Absolute Rest ring does not include an accelerometer, so position information was unavailable.

The respiratory rate estimation model exhibits degraded accuracy at extreme breathing rates, particularly below 10 bpm where errors exceeded 20% of the reference value. This limitation stems from training set imbalance: only 7.7% of BIDMC windows have respiratory rates below 10 bpm. For clinical populations with bradypnea (e.g., sedated surgical patients or central sleep apnea episodes), the current model may require additional calibration or training data enrichment in the low-rate regime.

Finally, the respiratory rate biomarker approach (HV Load and Event Index) did not generalize, with 0% sensitivity for apnea detection on MESA. While this negative result motivated the pivot to direct morphological features, it also means that the pipeline does not exploit respiratory dynamics for apnea detection, which limits interpretability compared to methods that directly model respiratory events.

6.4 Future Work

Several directions emerge from this work for future investigation.

The most immediate priority is expanding the clinical validation to larger and more diverse cohorts. A prospective study with 50 or more subjects per pathological class would provide the statistical power needed to establish clinical reliability and compute meaningful sensitivity, specificity, and negative predictive values. Including mild and moderate OSA patients would test the pipeline’s ability to detect less severe forms of the condition on wearable-grade signals. The study should include participants across a range of ages, body compositions, and comorbidity profiles to assess demographic generalization. Particular attention should be paid to patients with comorbid conditions that affect PPG morphology, such as atrial fibrillation, heart failure, and peripheral vascular disease, as these conditions are common in the sleep apnea population and could confound the classification.

Architectural improvements could further narrow the remaining gap to SpO₂-based methods. Attention mechanisms and Transformer architectures [103] could be explored to capture longer-range temporal dependencies than the current LSTM layers. Representation pretraining on large unlabeled PPG databases (such as the Apple Heart Study or UK Biobank wrist-worn PPG data) could improve the quality of learned features before supervised fine-tuning on the smaller labeled datasets.

The domain adaptation results suggest that unsupervised methods are insufficient for PPG-based apnea detection, but recent advances in contrastive and adversarial domain adaptation have not yet been explored in this context. Methods that align latent representations across domains while preserving class-discriminative information could potentially reduce the need for labeled target-domain data. A practical near-term approach would be to collect a small calibration set (10–20 labeled recordings) for each new sensor and evaluate whether this is sufficient for the frozen backbone adaptation to converge.

Multi-signal fusion represents another direction. While the classical PPG+SpO₂ fusion provided no benefit over SpO₂ alone, deep learning fusion of raw PPG and SpO₂ waveforms could capture complementary temporal patterns that hand-crafted features miss. Modern wearable devices increasingly include multiple sensors (PPG, accelerometer, temperature), and jointly modeling these

signals could improve both detection accuracy and specificity. Accelerometer data in particular would enable position-dependent analysis and motion artifact rejection, both of which are important for wearable deployment.

Real-time deployment optimization is necessary for practical use. The current pipeline processes full-night recordings offline, which introduces a delay between data acquisition and result delivery. Adapting the inference to streaming operation with segment-by-segment processing, confidence estimation, and alert generation would enable real-time monitoring. This would require addressing the challenge of patient-level aggregation in a streaming context: the top-K strategy requires all windows to be available before computing the patient score, so an incremental approximation (e.g., maintaining a running sorted buffer of the top K% window probabilities) would need to be developed and validated against the offline results.

6.5 Key Remarks

This work shows that a single PPG channel across clinical and wearable datasets contains information for multi-level respiratory analysis: from basic rate estimation to binary screening, continuous severity estimation, and cross-sensor transfer learning. The combination of the CoTELMo architecture, top-K aggregation, two-stage regression, and frozen-backbone adaptation forms a pipeline that addresses the main gaps identified in the introduction, namely limited cross-dataset validation, underexplored transfer learning for PPG-based apnea detection, and limited phenotype evidence on single-wavelength wearable-ring PPG.

The results suggest a practical deployment pathway. For binary screening, the frozen backbone strategy enables rapid adaptation to new sensors with minimal supervision (approximately 14,500 trainable parameters, weak patient-level labels only), as evaluated on CFS (N=324). For phenotype classification, the ExtraTrees classifier on handcrafted morphological features and the CoTELMo feature-based classifier achieved complete OSA-CSA separation in a small pilot study (N=46, 12 pathological); these results are hypothesis-generating and require prospective validation. The failure of respiratory rate biomarkers to generalize for apnea detection reinforces the importance of learning directly from the raw PPG waveform rather than relying on intermediate physiological estimates that may not capture the full complexity of sleep-disordered breathing.

While the small clinical sample sizes require cautious interpretation, the consistency of results across multiple datasets, methods, and evaluation frameworks provides a foundation for future clinical validation at scale. The key next steps are expanding the validation cohort (at least 50 subjects per pathological class), including mild-to-moderate OSA patients, and evaluating the pipeline on a broader range of consumer wearable devices to establish the generality of the frozen backbone adaptation approach.

References

- [1] M. Walker, *Why We Sleep: Unlocking the Power of Sleep and Dreams*. Scribner, 2017.
- [2] G. Medic, M. Wille, and M. E. H. Hemels, “Short- and long-term health consequences of sleep disruption,” *Nature and Science of Sleep*, vol. 9, pp. 151–161, 2017.
- [3] A. V. Benjafield *et al.*, “Estimation of the global prevalence and burden of obstructive sleep apnoea: A literature-based analysis,” *Lancet Respiratory Medicine*, vol. 7, no. 8, pp. 687–698, 2019.
- [4] Y. Yeghiazarians *et al.*, “Obstructive sleep apnea and cardiovascular disease: A scientific statement from the American Heart Association,” *Circulation*, vol. 144, no. 3, 2021.
- [5] S. Guo *et al.*, “Relationship between sleep apnea severity and mild cognitive impairment in people with obstructive sleep apnea,” *BMC Psychiatry*, vol. 25, no. 1, 2025.
- [6] S. Tregear, J. Reston, K. Schoelles, and B. Phillips, “Obstructive sleep apnea and risk of motor vehicle crash: systematic review and meta-analysis,” *Journal of Clinical Sleep Medicine*, vol. 5, no. 6, pp. 573–581, 2009.
- [7] American Academy of Sleep Medicine, “Hidden health crisis costing America billions,” Frost & Sullivan, Tech. Rep., 2016, commissioned by AASM. [Online]. Available: <https://aasm.org/resources/pdf/sleep-apnea-economic-crisis.pdf>
- [8] T. Young, L. Evans, L. Finn, and M. Palta, “Estimation of the clinically diagnosed proportion of sleep apnea syndrome in middle-aged men and women,” *Sleep*, vol. 20, no. 9, pp. 705–706, 1997.
- [9] A. N. Natsky *et al.*, “Economic evaluation of diagnostic sleep studies for obstructive sleep apnoea in the adult population: a systematic review,” *Sleep Medicine Reviews*, vol. 62, p. 101608, 2022.

- [10] N. A. Collop *et al.*, “Clinical guidelines for the use of unattended portable monitors in the diagnosis of obstructive sleep apnea in adult patients,” *Journal of Clinical Sleep Medicine*, vol. 3, no. 7, pp. 737–747, 2007.
- [11] P. H. Charlton, P. A. Kyriacou, J. Mant, V. Marozas, P. Chowienczyk, and J. Alastruey, “Wearable photoplethysmography for cardiovascular monitoring,” *Proceedings of the IEEE*, vol. 110, no. 3, pp. 355–381, 2022.
- [12] P. Dehkordi, A. Garde, B. Molavi, J. M. Ansermino, and G. A. Dumont, “Extracting instantaneous respiratory rate from multiple photoplethysmogram respiratory-induced variations,” *Frontiers in Physiology*, vol. 9, p. 948, 2018.
- [13] Y. Lin, X. Song, Y. Zhao, C. Zhang, and X. Ding, “Continuous respiratory rate monitoring through temporal fusion of ECG and PPG signals,” *PLOS ONE*, vol. 20, no. 6, p. e0325307, 2025.
- [14] P. H. Charlton *et al.*, “Breathing rate estimation from the electrocardiogram and photoplethysmogram: A review,” *IEEE Reviews in Biomedical Engineering*, vol. 11, pp. 2–20, 2018.
- [15] A. K. Kumar, M. Ritam, L. Han, S. Guo, and R. Chandra, “Deep learning for predicting respiratory rate from biosignals,” *Computers in Biology and Medicine*, vol. 144, p. 105338, 2022.
- [16] D. J. Eckert, A. S. Jordan, P. Merchia, and A. Malhotra, “Central sleep apnea: Pathophysiology and treatment,” *Chest*, vol. 131, no. 2, pp. 595–607, 2007.
- [17] R. N. Aurora, S. Chowdhuri, K. Ramar *et al.*, “The treatment of central sleep apnea syndromes in adults: Practice parameters with an evidence-based literature review and meta-analyses,” *Sleep*, vol. 35, no. 1, pp. 17–40, 2012.
- [18] F. Massie, S. Vits, A. Khachatryan, B. V. Pee, J. Verbraecken, and J. Bergmann, “Central sleep apnea detection by means of finger photoplethysmography,” *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 11, pp. 126–136, 2023.

- [19] Q. Zhao *et al.*, “Predicting respiratory rate from electrocardiogram and photoplethysmogram using a transformer-based model,” *Bioengineering*, vol. 10, no. 9, p. 1024, 2023.
- [20] M. Radha, P. Fonseca, A. Moreau, M. Ross, A. Cerny, P. Anderer, X. Long, and R. M. Aarts, “A deep transfer learning approach for wearable sleep stage classification with photoplethysmography,” *npj Digital Medicine*, vol. 4, p. 135, 2021.
- [21] Q. Li, Q. Li, A. S. Cakmak, G. D. Poian, D. L. Bliwise, V. Vaccarino, A. J. Shah, and G. D. Clifford, “Transfer learning from ECG to PPG for improved sleep staging from wrist-worn wearables,” *Physiological Measurement*, vol. 42, no. 4, p. 044004, 2021.
- [22] R. Marinheiro, L. Parreira, P. Amador, D. Mesquita, J. Farinha, M. Fonseca, T. Duarte, C. Lopes, A. Fernandes, and R. Caria, “Ventricular arrhythmias in patients with obstructive sleep apnea,” *Current Cardiology Reviews*, vol. 15, no. 1, pp. 64–74, 2019.
- [23] J. Allen, “Photoplethysmography and its application in clinical physiological measurement,” *Physiological Measurement*, vol. 28, no. 3, pp. R1–R39, 2007.
- [24] T. Tamura, Y. Maeda, M. Sekine, and M. Yoshida, “Wearable photoplethysmographic sensors—past and present,” *Electronics*, vol. 3, no. 2, pp. 282–302, 2014.
- [25] P. S. Addison *et al.*, “Developing an algorithm for pulse oximetry derived respiratory rate (RRoxi): A healthy volunteer study,” *Journal of Clinical Monitoring and Computing*, vol. 26, no. 1, pp. 45–51, 2012.
- [26] L. M. Nilsson, “Respiration signals from photoplethysmography,” *Anesthesia and Analgesia*, vol. 117, no. 4, pp. 859–865, 2013.
- [27] D. J. Meredith *et al.*, “Photoplethysmographic derivation of respiratory rate: a review of relevant physiology,” *Journal of Medical Engineering & Technology*, vol. 36, no. 1, pp. 1–7, 2012.
- [28] R. B. Berry *et al.*, “Rules for scoring respiratory events in sleep: Update of the 2007 AASM manual for the scoring of sleep and associated events,” *Journal of Clinical Sleep Medicine*,

- vol. 8, no. 5, pp. 597–619, 2012.
- [29] American Academy of Sleep Medicine Task Force, “Sleep-related breathing disorders in adults: Recommendations for syndrome definition and measurement techniques in clinical research,” *Sleep*, vol. 22, no. 5, pp. 667–689, 1999.
 - [30] L. J. Epstein, D. Kristo, P. J. Strollo *et al.*, “Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults,” *Journal of Clinical Sleep Medicine*, vol. 5, no. 3, pp. 263–276, 2009.
 - [31] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
 - [32] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
 - [33] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015, pp. 448–456.
 - [34] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
 - [35] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
 - [36] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
 - [37] D. Alves, F. Mendonça, S. S. Mostafa, and F. Morgado-Dias, “Time-series embeddings from language models: A tool for wind direction nowcasting,” *Journal of Meteorological Research*, vol. 38, pp. 558–569, 2024.

- [38] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [39] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.
- [40] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.
- [41] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *The Lancet*, vol. 327, no. 8476, pp. 307–310, 1986.
- [42] J. Cohen, "Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit," *Psychological Bulletin*, vol. 70, no. 4, pp. 213–220, 1968.
- [43] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.
- [44] J. Behar, A. Roebuck, J. S. Domingos, E. Geder, and G. D. Clifford, "A review of current sleep screening applications for smartphones," *Physiological Measurement*, vol. 34, no. 7, pp. R29–R46, 2013.
- [45] W. Karlen *et al.*, "Multiparameter respiratory rate estimation from the photoplethysmogram," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 7, pp. 1946–1953, 2013.
- [46] M. A. Pimentel *et al.*, "Toward a robust estimation of respiratory rate from pulse oximeters," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 8, pp. 1914–1923, 2017.
- [47] C. Orphanidou *et al.*, "Signal-quality indices for the electrocardiogram and photoplethysmogram: Derivation and applications to wireless monitoring," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 3, pp. 832–838, 2015.
- [48] R. Lazazzera and G. Carrault, "Breathing rate estimation methods from PPG signals, on CAPNOBASE database," in *2020 Computing in Cardiology*, 2020, pp. 1–4.

- [49] P. H. Charlton *et al.*, “An assessment of algorithms to estimate respiratory rate from the electrocardiogram and photoplethysmogram,” *Physiological Measurement*, vol. 37, no. 4, pp. 610–626, 2016.
- [50] D. Bian, P. Mehta, and N. Selvaraj, “Respiratory rate estimation using PPG: A deep learning approach,” in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020, pp. 5948–5952.
- [51] V. Ravichandran *et al.*, “RespNet: A deep learning model for extraction of respiration from photoplethysmogram,” in *2019 41st Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2019, pp. 5556–5559.
- [52] P. Osathitporn *et al.*, “RRWaveNet: A compact end-to-end multiscale residual CNN for robust PPG respiratory rate estimation,” *IEEE Internet of Things Journal*, vol. 10, no. 18, pp. 15 943–15 952, 2023.
- [53] M. H. Chowdhury, M. N. I. Shuzan, M. E. H. Chowdhury, M. B. I. Reaz, S. Mahmud, N. A. Emadi, M. A. Ayari, S. H. M. Ali, A. A. A. Bakar, and S. M. Rahman, “Lightweight end-to-end deep learning solution for estimating the respiration rate from photoplethysmogram signal,” *Bioengineering*, vol. 9, no. 10, p. 558, 2022.
- [54] M. N. I. Shuzan, M. H. Chowdhury, S. B. Alam, M. B. I. Reaz, M. S. Khan, M. Murugappan, and M. E. H. Chowdhury, “PPG2RespNet: A deep learning model for respiratory signal synthesis and monitoring from photoplethysmography signal,” *Physical and Engineering Sciences in Medicine*, vol. 47, no. 4, pp. 1705–1722, 2024.
- [55] C. S. Hwang, Y. H. Kim, J. K. Hyun, J. H. Kim, S. R. Lee, C. M. Kim, J. W. Nam, and E. Y. Kim, “Evaluation of the photoplethysmogram-based deep learning model for continuous respiratory rate estimation in surgical intensive care unit,” *Bioengineering*, vol. 10, no. 10, p. 1222, 2023.
- [56] H. J. Davies and D. P. Mandic, “Rapid extraction of respiratory waveforms from photoplethysmography: A deep corr-encoder approach,” *Biomedical Signal Processing and Control*, vol. 85, p. 104992, 2023.

- [57] H. Liu, J. Allen, D. Zheng, and F. Chen, “Recent development of respiratory rate measurement technologies,” *Physiological Measurement*, vol. 40, no. 7, p. 07TR01, 2019.
- [58] J. Chen, K. Sun, Y. Sun, and X. Li, “Signal quality assessment of PPG signals using STFT time-frequency spectra and deep learning approaches,” in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021, pp. 1153–1156.
- [59] C. Guilleminault, S. Connolly, R. Winkle, K. Melvin, and A. Tilkian, “Cyclical variation of the heart rate in sleep apnoea syndrome: mechanisms and usefulness of 24 h electrocardiography as a screening technique,” *The Lancet*, vol. 323, no. 8369, pp. 126–131, 1984.
- [60] E. Urtnasan, J.-U. Park, E.-Y. Joo, and K.-J. Lee, “Automated detection of obstructive sleep apnea events from a single-lead electrocardiogram using a convolutional neural network,” *Journal of Medical Systems*, vol. 42, no. 6, p. 104, 2018.
- [61] M. Bahrami and M. Forouzanfar, “Sleep apnea detection from single-lead ECG: a comprehensive analysis of machine learning and deep learning algorithms,” *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022.
- [62] F. Bozkurt, M. K. Uçar, M. R. Bozkurt, and C. Bilgin, “Detection of abnormal respiratory events with single channel ECG and hybrid machine learning model in patients with obstructive sleep apnea,” *IRBM*, vol. 41, no. 5, pp. 241–251, 2020.
- [63] S. S. Mostafa, F. Mendonça, A. G. Ravelo-García, and F. Morgado-Dias, “A systematic review of detecting sleep apnea using deep learning,” *Sensors*, vol. 19, no. 22, p. 4934, 2019.
- [64] M. Deviaene, D. Testelmans, B. Buyse, P. Borzee, S. V. Huffel, and C. Varon, “Automatic screening of sleep apnea patients based on the SpO₂ signal,” *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 607–617, 2019.
- [65] G. C. Gutiérrez-Tobal, D. Álvarez, A. Crespo, F. del Campo, and R. Hornero, “Evaluation of machine-learning approaches to estimate sleep apnea severity from at-home oximetry recordings,” *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 882–892, 2019.

- [66] A. Leino, S. Nikkonen, S. Kainulainen, H. Korkalainen, J. Töyräs, S. Myllymaa, T. Leppänen, S. Ylä-Herttuala, S. Western-Punnonen, A. Muraja-Murro, P. Jäkälä, E. Mervaala, and K. Myllymaa, “Neural network analysis of nocturnal SpO₂ signal enables easy screening of sleep apnea in patients with acute cerebrovascular disease,” *Sleep Medicine*, vol. 79, pp. 71–78, 2021.
- [67] J. Levy *et al.*, “Deep learning for obstructive sleep apnea diagnosis based on single channel oximetry,” *Nature Communications*, vol. 14, p. 4881, 2023.
- [68] J. Hayano *et al.*, “Quantitative detection of sleep apnea with wearable watch device,” *PLOS ONE*, vol. 15, no. 11, p. e0237279, 2020.
- [69] X. Jiang, Y. Ren, H. Wu, Y. Li, and F. Liu, “Convolutional neural network based on photoplethysmography signals for sleep apnea syndrome detection,” *Frontiers in Neuroscience*, vol. 17, p. 1222715, 2023.
- [70] E. S. J. Jothi, J. Anitha, and D. J. Hemanth, “A photoplethysmography-based diagnostic support system for obstructive sleep apnea using deep learning approaches,” *Computers & Electrical Engineering*, vol. 102, p. 108279, 2022.
- [71] P. Fonseca, N. den Teuling, X. Long, and R. M. Aarts, “Cardiorespiratory sleep stage detection using conditional random fields,” *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 4, pp. 956–966, 2017.
- [72] G. B. Papini, P. Fonseca, M. M. van Gilst, J. P. van Dijk, S. Overeem, J. W. M. Bergmans, and R. Vullings, “Wearable monitoring of sleep-disordered breathing: estimation of the apnea-hypopnea index using wrist-worn reflective photoplethysmography,” *Scientific Reports*, vol. 10, p. 13512, 2020.
- [73] K. Wei, L. Zou, G. Liu, and C. Wang, “MS-Net: Sleep apnea detection in PPG using multi-scale block and shadow module one-dimensional convolutional neural network,” *Computers in Biology and Medicine*, vol. 155, p. 106469, 2023.

- [74] H. Korkalainen, J. Aakko, B. Duce, S. Kainulainen, A. Leino, S. Nikkonen, I. O. Afara, S. Myllymaa, J. Töyräs, and T. Leppänen, “Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea,” *Sleep*, vol. 43, no. 11, p. zsaa098, 2020.
- [75] F. Mendonça, S. S. Mostafa, F. Morgado-Dias, and A. G. Ravelo-García, “An oximetry based wireless device for sleep apnea detection,” *Sensors*, vol. 20, no. 3, p. 888, 2020.
- [76] R. Casal, L. E. D. Persia, and G. Schlotthauer, “Classifying sleep-wake stages through recurrent neural networks using pulse oximetry signals,” *Biomedical Signal Processing and Control*, vol. 63, p. 102195, 2021.
- [77] M. Olsen, J. M. Zeitzer, R. Nakase-Richardson, V. H. Musgrave, H. B. D. Sorensen, E. Mignot, and P. J. Jennum, “A deep transfer learning approach for sleep stage classification and sleep apnea detection using wrist-worn consumer sleep technologies,” *IEEE Transactions on Biomedical Engineering*, vol. 71, no. 8, pp. 2506–2517, 2024.
- [78] I. Perez-Pozuelo, B. Zhai, J. Palotti, R. Mall, M. Aupetit, J. M. Garcia-Gomez, S. Taheri, Y. Guan, and L. Fernandez-Luque, “The future of sleep health: a data-driven revolution in sleep science and medicine,” *npj Digital Medicine*, vol. 3, p. 42, 2020.
- [79] G. H. Tison, J. M. Sanchez, B. Ballinger, A. Singh, J. E. Olgin, M. J. Pletcher, E. Vittinghoff, E. S. Lee, S. M. Fan, R. A. Gladstone, C. D. Mikell, N. Sber, E. E. Yeh, and A. C. Tseng, “Passive detection of atrial fibrillation using a commercially available smartwatch,” *JAMA Cardiology*, vol. 3, no. 5, pp. 409–416, 2018.
- [80] Z. Strumpf, W. Gu, C.-W. Tsai, P.-L. Chen, E. Yeh, L. Leung, C. Cheung, I.-C. Wu, K. P. Strohl, T. Tsai, R. J. Folz, and A. A. Chiang, “Belun Ring (Belun Sleep System BLS-100): Deep learning-facilitated wearable enables obstructive sleep apnea detection, apnea severity categorization, and sleep stage classification in patients suspected of obstructive sleep apnea,” *Sleep Health*, vol. 9, no. 4, pp. 430–440, 2023.
- [81] R. Lazazzera, Y. Belhaj, and G. Carrault, “A new wearable device for blood pressure estimation using photoplethysmogram,” *Sensors*, vol. 19, no. 11, p. 2557, 2019.

- [82] Y. Wang *et al.*, “A dual-modal wearable pulse detection system integrated with deep learning for high-accuracy and low-power sleep apnea monitoring,” *Advanced Science*, p. 2501750, 2025.
- [83] A. Abd-Alrazaq, H. Aslam, R. AlSaad, M. Alsahli, A. Ahmed, R. Damseh, S. Aziz, and J. Sheikh, “Detection of sleep apnea using wearable AI: Systematic review and meta-analysis,” *Journal of Medical Internet Research*, vol. 26, p. e58187, 2024.
- [84] A. Osa-Sanchez, J. R.-M. de Soria, A. Mendez-Zorrilla, I. O. Ruiz, and B. Garcia-Zapirain, “Wearable sensors and artificial intelligence for sleep apnea detection: A systematic review,” *Journal of Medical Systems*, vol. 49, p. 66, 2025.
- [85] F. Vaquerizo-Villar, D. Álvarez, L. Kheirandish-Gozal, G. C. Gutiérrez-Tobal, V. Barroso-García, E. Santamaria-Vázquez, D. Gozal, and R. Hornero, “Automatic sleep staging in children with sleep apnea using photoplethysmography and convolutional neural networks,” in *Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2021, pp. 216–219.
- [86] V. Birrer, M. Elgendi, O. Lambercy, and C. Menon, “Evaluating reliability in wearable devices for sleep staging,” *npj Digital Medicine*, vol. 7, p. 74, 2024.
- [87] B. Sun and K. Saenko, “Deep CORAL: Correlation alignment for deep domain adaptation,” in *European Conference on Computer Vision (ECCV) Workshops*, 2016, pp. 443–450.
- [88] M. Jafari *et al.*, “Application of transfer learning for biomedical signals: a comprehensive review of the last decade (2014–2024),” *Information Fusion*, vol. 118, p. 102982, 2025.
- [89] A. L. Goldberger *et al.*, “PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals,” *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
- [90] X. Chen *et al.*, “Racial/ethnic differences in sleep disturbances: The Multi-Ethnic Study of Atherosclerosis (MESA),” *Sleep*, vol. 38, no. 6, pp. 877–888, 2015.

- [91] G.-Q. Zhang *et al.*, “The National Sleep Research Resource: Towards a sleep data commons,” *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1351–1358, 2018.
- [92] S. Redline, P. V. Tishler, T. D. Tosteson *et al.*, “The familial aggregation of obstructive sleep apnea,” *American Journal of Respiratory and Critical Care Medicine*, vol. 151, no. 3, pp. 682–687, 1995.
- [93] M. Elgendi, “Optimal signal quality index for photoplethysmogram signals,” *Bioengineering*, vol. 3, no. 4, p. 21, 2016.
- [94] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.
- [95] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A next-generation hyperparameter optimization framework,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2623–2631.
- [96] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2015.
- [97] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995.
- [98] Y. Li, N. Wang, J. Shi, X. Hou, and J. Liu, “Adaptive batch normalization for practical domain adaptation,” *Pattern Recognition*, vol. 80, pp. 109–117, 2018.
- [99] B. Sun, J. Feng, and K. Saenko, “Return of frustratingly easy domain adaptation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.
- [100] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [101] R. S. Rosenberg and S. V. Hout, “The American Academy of Sleep Medicine inter-scorer reliability program: Respiratory events,” *Journal of Clinical Sleep Medicine*, vol. 10, no. 4, pp. 447–454, 2014.

- [102] S. S. Mostafa, F. Mendonça, F. Morgado-Dias, and A. Ravelo-García, “SpO₂ based sleep apnea detection using deep learning,” in *2017 IEEE 21st International Conference on Intelligent Engineering Systems (INES)*, 2017, pp. 91–96.
- [103] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.