Multi-Task Learning for Automated Cardiac Event Detection in Sleep Polysomnography

Foivos N Papathanasiou[1], Jing Han[1,2], Cecilia Mascolo[1]

[1] University of Cambridge, Cambridge, UK
[2] Hunan University, Changsha, China

Abstract

Sleep-disordered breathing is a failure of coupled cardiorespiratory regulation: repeated breathing disturbances destabilise cardiac control, with cardiovascular consequences. Polysomnography, the reference standard, is scored manually, which is slow and imperfectly reproducible. Automated detectors model the cardiac and respiratory event families separately, and none detects both jointly despite their physiological coupling and the success of multi-task learning (MTL) across health domains. In PSG-Audio, which annotates both families, we find statistically significant ordered coupling, respiratory events consistently preceding specific cardiac responses. To test whether MTL can exploit it, we develop CREST, to the best of our knowledge the first dense per-second detector jointly modelling cardiac and respiratory PSG events. Under patient-independent cross-validation it reaches macro-AUPRC 0.375 [0.356, 0.394] (respiratory 0.576, cardiac 0.124). While we provide evidence that cardiorespiratory coupling is measurable in sleep apnoea, we also show this correlation remains underutilised, a gap that should be addressed to advance automated sleep monitoring.

1. Introduction

Sleep is an active, tightly regulated physiological state rather than a period of rest. Across the night, breathing, heart rate, blood pressure and autonomic tone vary with sleep stage, reflecting coordinated regulation of the respiratory and cardiac systems [1]. This coupling becomes clinically important in sleep-disordered breathing, where repeated disturbances of ventilation destabilise oxygenation, arousal and cardiovascular control. Obstructive sleep apnoea, the most common form, alone affects an estimated 34% of men and 17% of women and remains largely undiagnosed [2]. Recurrent hypoxaemia and arousals contribute to cardiovascular and metabolic disease, including hypertension, arrhythmia and heart failure [2, 3].

Polysomnography (PSG), the overnight laboratory recording of brain, cardiac, respiratory and oxygenation signals, is the clinical reference standard for diagnosis, but its events must be scored manually, which is slow and imperfectly reproducible. Machine learning methods for automatic detection have been developed for many PSG tasks, including sleep staging, respiratory event detection, and arousal and limb-movement detection. Outputs range from epoch-level labels to dense per-time-point predictions.

Multi-task learning (MTL) has been effective across health domains [4], and the coupling between cardiac and respiratory sleep events makes them a natural pairing for it. Existing detectors nonetheless treat the two families separately: no PSG-signal detector produces dense per-second predictions spanning both the cardiac and respiratory families. The nearest multi-family detectors, MSED [5] and AttenCRF-U [6], reach event-level and dense output respectively but cover respiratory, limb-movement and arousal events, not cardiac ones, which are absent from the corpora these detectors train on. A substantial body of work learns from the cardiac signal, but predicts targets other than scored cardiac events.

Using PSG-Audio [7], a polysomnography corpus annotated with both cardiac and respiratory events, we test whether the two families are coupled in its annotations and whether MTL can exploit that coupling. This work makes three contributions: a demonstration of statistically significant ordered coupling between respiratory disturbances and downstream cardiac responses; the introduction of CREST (Cardio-Respiratory Event detection via Stream-Time attention), the first dense per-second detector jointly modelling cardiac and respiratory PSG events; and the finding that joint supervision of both families does not on its own improve cardiac event detection, with an occlusion analysis of the channels each detector relies on.

2. Cardiorespiratory Coupling

The expectation from physiology is an autonomic cascade: a respiratory event produces an arousal and a sympathetic surge [3], so respiratory events should be followed

by cardiac ones, long recognised on the electrocardiogram as progressive bradycardia through the apnoea and abrupt tachycardia on its resolution [8].

Methodological approach. Every experiment uses PSG-Audio [7], fixed at 193 recordings (868 h, 104,093 thirty-second windows). To quantify the coupling we applied event coincidence analysis [9] to the 139 recordings carrying an eligible event pair (32–106 contributing to each pair). It measures how often events of one type fall within a lag window ΔT of another, swept over nine values from 5 to 120 s, against a chance rate estimated here from circular-shift surrogates (mean of 1000 per patient-pair). The per-patient excess is tested by a Wilcoxon signed-rank test under Benjamini–Hochberg FDR correction ($\alpha = 0.05$); population-consistency requires a patient-level bootstrap interval above zero. Each pair's rate ratio, observed over chance, is quoted at the significant lag of largest median excess.

A coincidence is counted when a cardiac event begins within ΔT of the preceding respiratory event's end: the leading event is timed from its offset, the following one from its onset. Apnoeas are timestamped at airway-closure onset, but the cardiac response marks the apnoea's termination [8], so that is the interval the physiology spans. Anchoring both events at onset instead inverts the short-lag direction, an artefact of apnoea clustering rather than a physiological ordering, since a cardiac response cannot precede the apnoea it follows; the forward cascade remains FDR-significant under both anchorings.

Forward-coupling results. Under offset anchoring a coherent forward apnoea → cardiac cascade emerges as the dominant short-range coupling (Fig. 1). Its core is an apnoea/hypopnoea → PttDrop band whose excess peaks at 10–20 s and is largest for MixedApnea → PttDrop at 10 s (rate ratio 3.73); it stays FDR-significant and population-consistent to 120 s, the longest lag swept. A weaker, slower MixedApnea → Bradycardia tier, consistent with vagally mediated effects, reaches significance only from 20 s and population-consistency only at the longest lags. Short-lag couplings onto Tachycardia and LongRR also reach significance, strongest as ObstructiveApnea → Tachycardia (rate ratio 1.78 at 10 s), but are not population-consistent.

Reverse-coupling results. The reverse direction (cardiac leading respiratory) is markedly weaker than the forward cascade: no population-consistent rate ratio in that direction exceeds 1.64.

3. CREST

To exploit this coupling we develop CREST, a dense detector in two stages (Fig. 2): the OSF sleep foundation model [10] encodes each 30 s window on its own, and a cross-epoch model then mixes the resulting per-second tokens across neighbouring windows to predict densely at

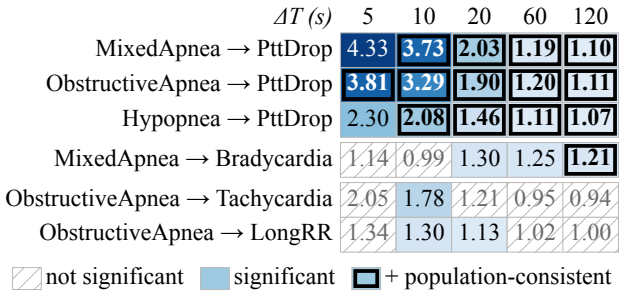

Figure 1. Offset-anchored coupling: rate ratio at five of nine lags. Respiratory leads cardiac; reverse not shown.

1 Hz. CREST trains roughly 5.54 M parameters on the frozen ∼85 M backbone.

Backbone. OSF is a 12-layer ViT-Base that tokenises each window into one patch token per second for each of its three channel-groups: cardiac (ECG, EMG), respiratory (effort, nasal pressure, snore) and brain (EEG, EOG). It bundles the ECG with the three EMG channels into one group, which dilutes it. CREST zeroes those EMG channels before tokenisation. Its CLS token is dropped and the per-second patch tokens kept. The encoder is adapted by LoRA [11] ($r = 8$); the tokeniser and positional embeddings stay frozen.

Signal routing. Two side-paths enter the cross-epoch model as additional per-second streams. The first carries SpO_2: a Linear($1 \to 32$) projection of its native 1 Hz trace. The second is a learned ECG side-path that passes clean cardiac signal around that bundling. The ECG side-path is a dual-resolution convolutional stem: a six-layer fine branch ($k = 7$, $s = 2$) for the fast rate changes of Tachycardia and a coarse convolution ($k = s = 64$) for the slow changes of Bradycardia, both reaching the 1 Hz grid.

Window mixing. The cross-epoch model is a stream-time Transformer. It keeps the five streams as separate tokens at width 384 and runs full self-attention across the stream-time grid. The model spans five consecutive 30 s windows (a 150 s context) and predicts the central one, so each predicted second carries at least 60 s of context on either side. That reach covers the apnoea/hypopnoea → PttDrop band's 10–20 s excess peak and the lags at which the slower Bradycardia tier reaches significance.

Output head. Many events are shorter than the standard 30 s epoch, so an epoch-level label would dilute the short markers; CREST's dense output head is therefore a single Linear($d \to K$): K per-second logits, independent sigmoids at inference. Here $K = 9$, the cardiac and respiratory classes of Section 4.

Table 1. Joint nine-class detector vs. single-family baselines. prev. is chance AUPRC; Sens., Spec. and Prec. are per-second at per-fold inner-validation thresholds; ΔAUPRC is joint minus same-family baseline; p is Benjamini–Hochberg-corrected.

			operating point				
Class	prev.	AUPRC [95% CI]	Sens.	Spec.	Prec.	ΔAUPRC [95% CI]	p
Respiratory family							
RelativeDesaturation	0.231	0.859 [0.837, 0.879]	0.802	0.920	0.751	−0.017 [−0.024, −0.010]	<0.001
ObstructiveApnea	0.163	0.803 [0.768, 0.834]	0.792	0.937	0.708	−0.021 [−0.031, −0.011]	<0.001
MixedApnea	0.044	0.582 [0.480, 0.664]	0.696	0.967	0.488	−0.051 [−0.082, −0.023]	<0.001
Hypopnea	0.052	0.323 [0.293, 0.354]	0.486	0.943	0.319	−0.048 [−0.063, −0.034]	<0.001
CentralApnea	0.009	0.312 [0.248, 0.379]	0.488	0.990	0.307	−0.023 [−0.083, +0.037]	0.405
Cardiac family							
LongRR	0.018	0.155 [0.110, 0.206]	0.274	0.982	0.218	+0.013 [−0.008, +0.033]	0.210
Tachycardia	0.011	0.149 [0.108, 0.200]	0.291	0.983	0.158	+0.013 [−0.020, +0.045]	0.432
Bradycardia	0.040	0.097 [0.051, 0.164]	0.311	0.924	0.145	+0.011 [−0.003, +0.028]	0.130
PttDrop	0.039	0.094 [0.053, 0.149]	0.273	0.862	0.074	−0.014 [−0.035, +0.005]	0.153
macro (9)	—	0.375 [0.356, 0.394]	—	—	—	—	—
respiratory (5)	—	0.576 [0.546, 0.602]	—	—	—	−0.032 [−0.048, −0.016]	<0.001
cardiac (4)	—	0.124 [0.097, 0.158]	—	—	—	+0.006 [−0.007, +0.018]	0.351
Pooled macro-AUROC (9 classes): 0.852 [0.835, 0.871]							

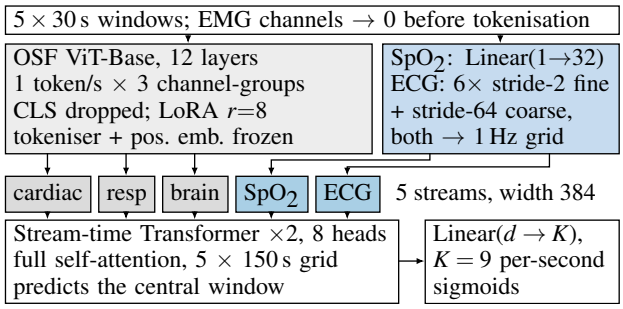

Figure 2. CREST in data-flow order; the five streams are OSF's three channel-groups plus two side-paths.

4. Experimental Setup

Classes and heads. Nine PSG-Audio event types [7] are used as supervised targets, divided into two physiological systems: four cardiac classes (Bradycardia, Tachycardia, LongRR and PttDrop) and five respiratory ones (ObstructiveApnea, CentralApnea, MixedApnea, Hypopnea and RelativeDesaturation, the SpO_2 marker grouped here with them). The joint-versus-single comparison uses three head configurations of one backbone and training recipe, differing in which families are supervised: the joint detector over all nine classes, and cardiac-only (four) and respiratory-only (five) single-family baselines.

Labelling. Between 72 and 100% of every class falls below one 30 s epoch, though durations span the fixed 2 s LongRR marker to multi-minute cardiac runs. Supervision is one multi-label vector per second; a second is labelled positive for a class whenever its half-open interval intersects an annotated event of that class.

Cross-validation. All reporting is patient-independent, over five folds with 38–40 test patients each. The folds are drawn by first-order iterative stratification, which balances a 193 × 17 covariate matrix across folds: class-presence indicators (9), AHI-severity bins (3), cardiac-burden quartiles (4) and a sex indicator (1).

Training. All variants share one recipe: positive-weighted binary cross-entropy over the dense 1 Hz targets ($N_{\text{neg}}/N_{\text{pos}}$ per class, per fold), averaged over seconds and classes; AdamW at 10^{-3} with linear warmup then cosine anneal; effective batch size 256; ten epochs, fixed, with no early stopping.

Scoring. Detection is scored by AUPRC per class, per family and overall, and by macro-AUROC, for all three detectors, pooled over patients with a 10,000-resample bootstrap confidence interval. Sensitivity, specificity and precision are given at the operating point.

5. Results and Discussion

Detection performance. Joint supervision of both families does not improve cardiac event detection. Against the single-family baselines no cardiac ΔAUPRC is significant, the cardiac macro standing at +0.006 with an interval spanning zero, while the respiratory family pays a significant −0.032, all five of its classes negative and four significant after correction (Table 1), a negative transfer consistent with the two families competing for one shared encoder.

The gain is coverage at close to single-family accuracy: one dense nine-class detector spans both families in a single 1 Hz pass, reaching a macro-AUPRC of 0.375. The cardiac classes are weak in a specific way: their sensitivity outruns their precision, so their errors are false alarms rather than misses.

Channel reliance. Occluding input streams at inference

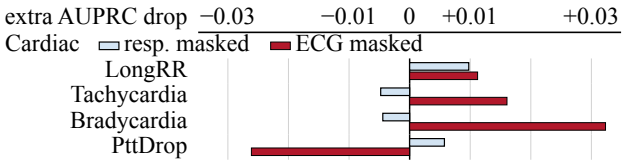

Figure 3. Cardiac occlusion at inference: extra AUPRC lost by the joint detector against its same-family baseline. Five-fold means, no confidence intervals.

reads what each class relies on; Fig. 3 gives that reading on the cardiac classes, for the joint detector net of its same-family baseline, so a positive bar means the joint detector lost more. The respiratory mask sits near zero on all four cardiac classes: joint training barely moved their reliance on those channels against the cardiac-only baseline, though both detectors still lose accuracy on three of the four. What did move is ECG reliance: the joint detector loses more than the baseline under ECG removal on exactly the three cardiac classes whose ΔAUPRC is positive (Table 1), while PttDrop reverses on both measures. Respiratory supervision pushing the shared encoder to read ECG structure would surface exactly this way, since respiration reads its own channels at inference. The reading is speculative: greater reliance on a channel is not the same as a better representation, inference-time masking measures distribution-shift sensitivity as well as reliance, and generic multi-task regularisation is not excluded.

Temporal reach. We widened the cross-epoch context from 150 s to 210 and 270 s, the wider of which spans the longest measured coupling lag; yet macro-AUPRC did not move. Therefore, context length does not appear to be the limit, unless the architecture fails to use the extra context.

Sparsity. What limits the cardiac comparison is how rare the cardiac events are: each fold holds only a handful of patients carrying them, so the cardiac intervals are wide, and PttDrop, present in just over a quarter of recordings, most plausibly owes its negative estimate to that sparsity.

Cohort and backbone. Every result rests on one corpus, with no separate held-out test set, on a severely skewed, mostly male referral sample (88% severe apnoea, 74% male). OSF was pretrained to represent 30 s windows, not fine cardiac detail: it cannot isolate clean cardiac tokens (Section 3), and it resamples to 64 Hz. Above all, its self-supervised pretraining objective does not explicitly model the cardiorespiratory coupling this task turns on.

In conclusion, we have presented CREST, to the best of our knowledge the first detector of cardiac and respiratory PSG events jointly at dense per-second resolution, and established a reference baseline for this new task. Joint supervision of both families did not improve cardiac event detection on its own, which motivates designs that engage the coupling explicitly. We propose three, none tested here: a backbone pretrained from scratch on large sleep corpora with a cardiac-preserving objective and a target that models the coupling itself; lag-aware structure, since the coupling's timescale differs across event pairs (cf. Section 2) while the context here is one fixed symmetric window; and conditioning cardiac prediction on detected respiratory events.

Acknowledgments

This work was supported by ERC Project AURORA (101266067).

References

[1] Fink AM, Bronas UG, Calik MW. Autonomic regulation during sleep and wakefulness: a review with implications for defining the pathophysiology of neurological disorders. Clin Auton Res 2018;28(6):509–518.

[2] Javaheri S, Barbe F, Campos-Rodriguez F, et al. Sleep apnea: Types, mechanisms, and clinical cardiovascular consequences. J Am Coll Cardiol 2017;69(7):841–858.

[3] Dempsey JA, Veasey SC, Morgan BJ, et al. Pathophysiology of sleep apnea. Physiol Rev 2010;90(1):47–112.

[4] Zhao Y, Wang X, Che T, et al. Multi-task deep learning for medical image computing and analysis: A review. Comput Biol Med 2023;153:Art. no. 106496.

[5] Zahid AN, Jennum P, Mignot E, et al. MSED: A multimodal sleep event detection model for clinical sleep analysis. IEEE Trans Biomed Eng 2023;70(9):2508–2518.

[6] Li Q, Li K, Fu C, et al. AttenCRF-U: Joint detection of sleep-disordered breathing and leg movements in OSA patients. Bioengineering 2025;12(6):Art. no. 571.

[7] Korompili G, Amfilochiou A, Kokkalas L, et al. PSG-Audio, a scored polysomnography dataset with simultaneous audio recordings for sleep apnea studies. Sci Data 2021; 8(1):Art. no. 197.

[8] Guilleminault C, Winkle R, Connolly S, et al. Cyclical variation of the heart rate in sleep apnoea syndrome. Lancet 1984;323(8369):126–131.

[9] Donges JF, Schleussner CF, Siegmund JF, et al. Event coincidence analysis for quantifying statistical interrelationships between event time series. Eur Phys J Spec Top 2016; 225(3):471–487.

[10] Shuai Z, Xu Z, Yang D, et al. OSF: On pre-training and scaling of sleep foundation models. arXiv:2603.00190 [cs.LG], Feb. 2026.

[11] Hu EJ, Shen Y, Wallis P, et al. LoRA: Low-rank adaptation of large language models. Proc. Int. Conf. Learn. Represent. (ICLR), 2022. arXiv:2106.09685.

Address for correspondence:

Foivos N Papathanasiou
15 JJ Thomson Avenue, Cambridge, UK, CB3 0FD
fnp23@cam.ac.uk