CogPhys: Assessing Cognitive Load via Multimodal Remote and Contact-based Physiological Sensing

Anirudh Bindiganavale Harish[1*], **Peikun Guo**[1*], **Bhargav Ghanekar**[1†], **Diya Gupta**[1†],
Akilesh Rajavenkatanarayanan[2], Manoj Kumar Sharma[2], Maureen August[2],
Akane Sano[1], Ashok Veeraraghavan[1]

[1]Rice University, [2]General Motors

{anirudhbh, Peikun.Guo, bhargav.ghanekar, Diya.Gupta, akane.sano, vashok}@rice.edu
{akilesh.rajavenkatanarayanan, maureen.short}@gm.com
manojsharma.iitd@gmail.com

* - contributed equally to the project
† - contributed equally to the project

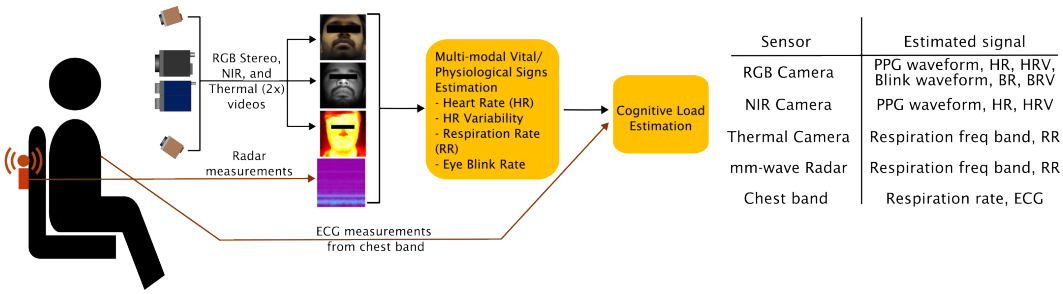


Figure 1: We propose **CogPhys**, a dataset consisting of multimodal recordings of seated participants while performing tasks of varying cognitive loads. RGB Stereo, NIR, and two thermal videos are captured in conjunction with radar recordings to first estimate biosignals, such as plethysmographs (PPG), respiratory waveforms, and blink waveforms. The associated vital signs - heart rate (HR), heart rate variability (HRV), respiratory rate (RR), and physiological signals such as blink rate (BR), blink rate variability (BRV) - can be estimated from the previously extracted biosignals. Cognitive load, a higher-order physiological signal, is then estimated from the vital and physiological signs extracted in the previous stage, thus creating a pipeline for remote cognitive load estimation.

Abstract

Remote physiological sensing is an evolving area of research. As systems approach clinical precision, there is increasing focus on complex applications such as cognitive state estimation. Hence, there is a need for large datasets that facilitate research into complex downstream tasks such as remote cognitive load estimation. A first-of-its-kind, our paper introduces an open-source multimodal multi-vital sign dataset consisting of concurrent recordings from RGB, NIR (near-infrared), thermal, and RF (radio-frequency) sensors alongside contact-based physiological signals, such as pulse oximeter and chest bands, providing a benchmark for cognitive state assessment. By adopting a multimodal approach to remote health sensing, our dataset and its associated hardware system excel at modeling the complexities of cognitive load. Here, cognitive load is defined as the mental effort exerted during tasks such as reading, memorizing, and solving math problems. By using the

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

NASA-TLX survey, we set personalized thresholds for defining high/low cognitive levels, enabling a more reliable benchmark. Our benchmarking scheme bridges the gap between existing remote sensing strategies and cognitive load estimation techniques by using vital signs (such as photoplethysmography (PPG) and respiratory waveforms) and physiological signals (blink waveforms) as an intermediary. Through this paper, we focus on replacing the need for intrusive contact-based physiological measurements with more user-friendly remote sensors. Our benchmarking demonstrates that multimodal fusion significantly improves remote vital sign estimation, with our fusion model achieving $< 3\ BPM$ (beats per minute) error for vital sign estimation. For cognitive load classification, the combination of remote PPG, remote respiratory signals, and blink markers achieves 86.49% accuracy, approaching the performance of contact-based sensing (87.5%) and validating the feasibility of non-intrusive cognitive monitoring. Github Codebase: https://github.com/AnirudhBHarish/CogPhys

1 Introduction

Non-contact physiological monitoring is rapidly advancing, with applications extending beyond traditional healthcare into diverse domains [31, 40, 36]. Various remote sensors, including video, radars, and thermal cameras, are increasingly employed, often in multimodal setups, to estimate vital signs like heart rate (HR), respiratory rate (RR), and heart rate variability (HRV) [49, 39, 38]. While many datasets facilitate research in remote cardiovascular and respiratory assessment [5, 31, 38] (see Table 1), resources for higher-level physiological states, such as cognitive load, especially using diverse *remote* modalities, remain scarce [22]. This highlights a critical need for datasets enabling advanced remote sensing applications like non-intrusive cognitive load estimation, which is crucial in scenarios where contact-based sensors are impractical.

Cognitive load, defined as the mental effort required to perform a task, has been extensively studied in the context of human-computer interaction, workload assessment, and physiological monitoring. While self-reported metrics such as the NASA-TLX survey [15] allow for systematic assessments of cognitive load, they are inherently intrusive to the task being performed and lack the potential for real-time automation. To address this limitation, we explore using physiological markers as implicit indicators of cognitive workload. High cognitive load activates the sympathetic nervous system, which elevates HR and reduces HRV [3, 34]. Conversely, low cognitive load is associated with increased parasympathetic activity, which helps the body relax after periods of stress. Cognitively demanding tasks also accelerate respiration due to increased brain oxygen demands [14]. Additionally, behavioral markers like eye blink rate (BR) and blink rate variability (BRV) correlate with cognitive load [27]. These established physiological relationships enable the use of HR, RR, and blink patterns to distinguish cognitive load levels.

Thus, the accurate, robust measurement of physiological and vital signs could potentially pave the way for estimating abstract, higher-order physiological signals such as cognitive load. Contact-based sensing is not always feasible due to user burden. By combining the advances in remote sensing with the correlation between various vital signs and cognitive load, we can take remote physiological sensing a step further. Integrating remote monitoring systems can potentially enhance the applicability of systems for domains such as at-home rehabilitation, telemedicine, driver monitoring, etc.

To this end, we adopt a multimodal approach to remote sensing and cognitive load estimation. Signals are recorded from a sensor suite consisting of RGB, near-infrared (NIR), and Thermal cameras, and a radar, as participants perform tasks involving varying cognitive loads. Uniquely, our dataset introduces a total of five remote modalities, more than twice those present in existing datasets. This medley of data sources enables research into sophisticated and intelligent fusion strategies that improve robustness and mitigate bias [38], for estimating both conventional (e.g., photoplethysmography (PPG) waveform) as well as abstract (e.g., cognitive load) physiological signals. We split this task into 2 stages - the first being remote vital sign estimation, and the second, cognitive load prediction from the estimated vital signs. To summarize, our contributions are as follows:

1. We present a first-of-its-kind multi-modal dataset, called *CogPhys*, that records participants performing tasks with varying levels of cognitive load through cameras (RGB/NIR/Thermal), radar, pulse oximeter, and a wearable chest band.

2. We benchmark contemporary remote vitals sign estimation algorithms - both unimodal methods and multimodal sensors fusion algorithms - on our dataset. We accomplish this by estimating the biosignals and extracting vital signs, such as HR, HRV, and RR.
3. We then proceed to estimate and classify cognitive load by utilizing the various physiological features extracted from the biosignals and their corresponding biosignals. Further, we benchmark the performance of various machine learning (ML) and deep learning (DL) models on a combination of the vital sign modalities

2 Related Work

2.1 Remote Vital Sensing

Remote sensing of HR, HRV, and RR using cameras and radars has gained popularity [40, 33, 7]. Large-scale datasets for remote photoplethysmography (rPPG) and remote respiration (resp) have emerged across various settings: lab environments [5], NIR imaging [30], driving scenarios [31], and synthetic data [28, 40]. These datasets fostered model-based approaches including spatial Signal-to-Noise Ratio (SNR) maps [21, 7], sparse spectral methods [31, 30], light transport techniques [9, 39], and motion-based methods [4]. DL has advanced rPPG estimation through transformers [47, 46], mamba architectures [26, 51], and contrastive networks [35, 36]. Parallel research has explored respiratory signal estimation using thermal cameras and radio-frequency sensors [8], with novel data augmentation techniques for ultra-wideband (UWB) and frequency modulated continuous wave (FMCW) radars [49, 38].

Relying on a single modality for remote vital sensing often yields systems vulnerable to lighting changes and inaccuracies, potentially causing inequitable estimation errors. Several works have explored multimodal approaches, with datasets enabling advances in remote multimodal vital sensing [48, 38, 19, 30, 31]. This has led to various advanced algorithms benchmarked on both unimodal as well as multimodal data streams [36, 42]. rPPG-based methods are sensitive to face motions, and [24] attempted to reduce motion corruption by adaptively filtering the ballistocardiography (BCG) signal and rPPG signals. Similarly, [2] has extended the system's capabilities to include the fusion of remote ballistocardiography (rBCG) signals in addition to rPPG and contact-based BCG.

2.2 Cognitive State Estimation

ML has been increasingly leveraged to estimate cognitive workload from physiological signals. Contact-based approaches often use Electroencephalography (EEG) for direct brain activity measurement [43] or wearable sensors for extracting electrocardiography (ECG) and PPG signals to infer cognitive states via features such as HRV [13, 17]. Non-contact methods primarily involve cameras for eye-tracking (gaze, pupil dynamics) [20] or assessing behavioral indicators like facial expressions [17]. While remote sensing with modalities such as radar shows promise for vital signs, its direct use for cognitive state estimation is less explored.

3 Dataset Design

3.1 Captured Biosignals and Physiological Signals - Remote and Contact-based

PPG is a non-invasive optical method for measuring blood volume changes, providing vital signs like **HR** and **HRV**. It is typically recorded using pulse oximeters. Replacing these with RGB or NIR cameras enables **rPPG** by detecting skin color changes from blood flow. HR is measured in BPM and HRV in **milliseconds** (ms). A chest band sensor was also used for **ECG** recordings, but due to the need for skin contact, usable ECG data was only collected from a subset of participants.

Similarly, thermal cameras offer an optical method for sensing **breathing signals** and **RR** - measured in **respirations per minute** (RPM) - by detecting temperature changes near the nostrils during inhalation and exhalation. The radar's high-phase sensitivity can capture breathing-induced vibrations when placed in front of or behind the user. Capacitive sensors in chest bands provide baseline RR readings by tracking impedance changes caused by the expansion of the chest during breathing cycles.

Table 1: **List of the most popular datasets for remote vital sensing.** Each checkmark represents the use of one sensor, i.e., a double checkmark represents the use of 2 cameras.

	Recorded Signals				Ground Truth		
Dataset	**RGB**	**NIR**	**RF**	**Therm**	**HR**	**Resp**	**Cog Load**
UBFC [5]	✓				✓		
MMPD [37]	✓				✓		
VIPL [29]	✓				✓		
iBVP [19]	✓			✓	✓		
MMSE [48]	✓			✓	✓	✓	
UCLA-rPPG [40]	✓				✓		
SCAMPS [28]	✓				✓		
OOD-rPPG [6]	✓				✓		
MR-NIRP [31]	✓	✓			✓		
EquiPleth [38]	✓		✓		✓		
Ours (CogPhys)	✓✓	✓	✓	✓✓	✓	✓	✓

Sensor	**Description**
RGB	Zed2i Stereo
NIR	GS3U3-41C6NIRC
NIR 940nm LED	ThorLabs
LED Collimator	ThorLabs
940nm Filter	Edmund Optics
Thermal Rad	Boson 640
Thermal Non-Rad	Boson 640
Radar	AWR6843ISK
FPGA	DCA100EVM
Pulse Oximeter	CMS60C
Chest Band	BioHarness 3.0

Figure 2: CogPhys Multi-modal capture setup. We record signals from seated participants using various non-contact sensors (RGB Stereo camera, NIR camera, two thermal cameras, mm-wave radar) and contact wearables (chest band for ECG and respiratory signals, and finger pulse oximeter). The participants perform several tasks, such as sitting still, reading, performing arithmetic operations, etc., while the signals are recorded.

Alongside extracting rPPG waveforms, an RGB camera can also track eye landmarks to measure the vertical opening of the eye, referred to as the **"eye openness" (EO) signal** [32]. From this 1D signal, blink-related metrics such as **BR** and **BRV** can be derived by tracking the signal minima. Blinking, a semi-autonomic response, is governed by the central nervous system and has been previously linked to cognitive load [16, 1].

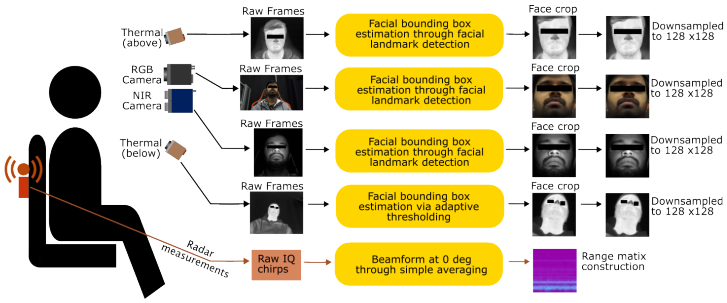
3.2 Experimental Setup used to Capture Data

Our sensing setup prioritizes non-contact vital sign monitoring using an array of remote sensors, including radar, RGB, NIR, and two thermal cameras. As shown in Figure 2, the RGB stereo camera, NIR camera, and their illuminators are mounted on the cockpit directly in front of the participant to best capture facial color changes. The NIR system is operated at $940\ nm$ using a bandpass filter following the findings of [30] on the reduced sensitivity of $940\ nm$ to sunlight, making it well-suited for outdoor use. The two thermal cameras are mounted above and below the cockpit to accurately capture temperature changes throughout the face and nostrils, respectively. In contrast, the radar is placed behind the seat within a plastic enclosure.

We collect baseline readings from a wearable chest band and a fingertip pulse oximeter. These serve two purposes: training/validating remote vital sign algorithms, and providing a reference signal for advanced tasks such as cognitive load prediction from physiological signals, thereby establishing upper bounds.

Tasks Performed in sequence (2 min. each)	Expected Cognitive Load
Sitting still	Low
Reading a paragraph	Medium
Rest period	N/A
Memorize 4x4 pattern; answer math questions	High
Rest period	N/A
Memorize 8-digit number; answer math questions	High

(A) Data acquisition procedure



(B) Data preprocessing steps

Figure 3: (A) Participants perform the illustrated tasks sequentially, with each task duration being 2 $mins$. The incoming raw data from all remote sensing modalities is preprocessed as shown in (B) before estimations of the PPG and Respiratory waveforms are made.

3.3 Data Acquisition Protocol

With our multimodal data collection setup firmly in place, we look to investigate the efficacy of remote vital sensing for cognitive load prediction. We define cognitive load as the mental effort that an individual exerts when performing a given task. To emulate diverse cognitive states, we curate a set of tasks to induce varying levels of cognitive load, which are summarized as:

1. **Still:** The participant sits still throughout the recording.
2. **Read:** The participant reads a set of independent, randomly generated paragraphs.
3. **Pattern:** Participants memorize a 4×4 binary grid, answer multiplication questions during the recording, then reproduce the grid upon completion of the recording.
4. **Number:** It is similar to "*pattern*", except participants are asked to memorize an 8-digit number and recite it at the end of the recording.

Each task is performed for 2 $mins$, with two resting periods — one between "*read*" and "*pattern*", and another between "*pattern*" and "*number*" — resulting in six 2 $mins$ recordings. The "rest" recordings allow participants to recover between trials and reduce the cumulative effects of the trials. The two "rest" trials are solely used for remote vitals sensing and not for cognitive load prediction. Participants also provided self-reported NASA-TLX ratings on mental demand, effort, and frustration after each of the 4 cognitive tasks, enabling a dual-perspective evaluation of cognitive workload.

We collect a large-scale dataset with 37 participants recruited for our study. This dataset, dubbed the *CogPhys* dataset, was collected in compliance with an Institutional Review Board (IRB) (approved by the Rice University Institutional Review Board, study number: IRB-FY2025-59). In addition to the sensor data, we also collect demographic details such as the use of glasses/contacts, cosmetics, age, gender, self-reported Fitzpatrick skin tone values, and height. Across our study, we capture ≈ 440 $mins$ video and radar recordings. For all our experiments, we partition the dataset into train, validation, and test sets with a split corresponding to 25/2/10 with no participant overlap. Dataset details such as demographic distribution, pre-processing algorithms (cropping, downsampling), storage size, and dimensionality are all elaborated in the Supplement.[1]

3.4 Label Curation for Cognitive Load

Cognitive load ground truth labels ("Low Load"/"High Load") for each participant-task instance were derived from self-reported NASA-TLX scores by applying participant-specific median thresholds. This approach was chosen to leverage subjective assessments while ensuring robustness against inter-participant reporting biases and preserving individual variance in workload perception.

The label generation process was as follows: First, for each participant, a composite workload score (0-80 range) was calculated for each of the four designated cognitive tasks. This composite

[1]To access this dataset, researchers can contact the authors and must execute a Data Use Agreement (DUA) due to the nature of the collected data.

score approach follows standard practices in cognitive workload research, where the NASA-TLX is designed as a multidimensional assessment tool [41, 44]. This score aggregated their ratings from four demand-relevant NASA-TLX subscales: Mental Demand, Temporal Demand, Effort, and Frustration. Second, the median of these composite scores, determined across a participant's set of cognitive tasks, established their personalized workload threshold. Finally, this individual threshold was used to binarize the cognitive state for each task instance. While cognitive load exists on a continuous spectrum, we adopt a binary classification approach for this first investigation of remote cognitive load sensing to establish feasibility before extending to multi-class or regression formulations in future work. The data-driven labeling resulted in labels that aligned with intended experimental task difficulties, visualized in Figure 4.

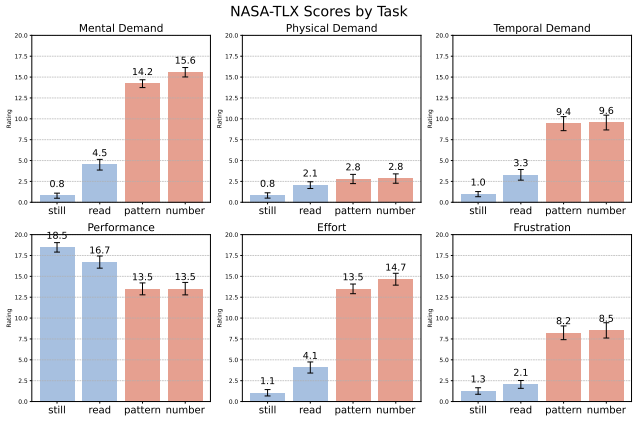


Figure 4: NASA-TLX questionnaire responses across different tasks. Participant average ratings with standard error bars for each of the six NASA-TLX dimensions across the four experimental tasks. Samples from {still,read}/{pattern,number} were labeled low/high cognitive load, respectively.

4 Estimating Remote Physiological Signals from a Sensor Stack

We now describe the deep-learning algorithms used to extract the vital and physiological signals from Section 3.1, focusing on methods with superior performance [23] and multimodal fusion approaches. Additional algorithmic baselines and implementation details are provided in the Supplement.

4.1 Remote Cardiac Monitoring

We employ the rPPG-Toolbox [23][2] and Contrast-Phys+ [36] code repositories[3] to benchmark the most significant rPPG baselines. Crucially, we include the *SNR Loss* from [38] and filter the estimated and ground truth waveforms prior to the Pearson loss. Further details on the baselines can be found in the supplement. Apart from the unimodal approaches, we train a fusion network.

Fusion Network: We employ a Siamese network with the Contrast-Phys+ backbone. That is, we share the weights of the first 2 convolutional blocks across the RGB and NIR videos, after which the deep features are added and passed through the remaining layers. Our pretraining strategy leverages the higher SNR of RGB signals: we first train a Contrast-Phys+ model on RGB frames, then use its weights to initialize and fine-tune a NIR model. Finally, we initialize the fusion network with the trained NIR model weights to prevent RGB dominance during multimodal training.

4.2 Respiratory Signal and Rate

The rPPG-Toolbox was adapted for RR estimation from thermal camera data. That is, we modify cutoff frequencies for frequency-based loss functions. We also downsample input videos and ground truth signals to 15 Hz to enable processing longer video segments at lower computational costs.

[2]https://github.com/ubicomplab/rPPG-Toolbox

[3]We port the code from the original repository to the rPPG-toolbox and add the SNR and Person loss function.

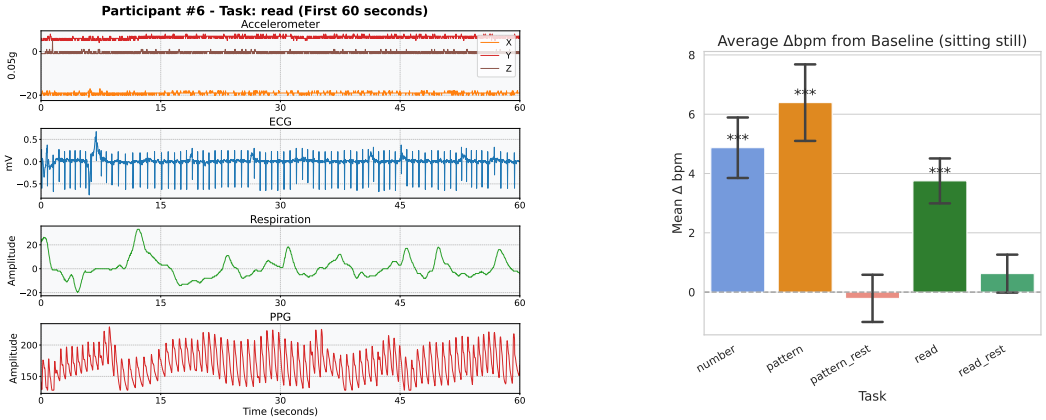


Figure 5: **Left:** Example of vital signs waveforms recorded by contact sensors. **Right:**Task-induced changes in HR (BPM) compared to the baseline condition. The bars represent mean differences. Standard error bars with asterisks "***" denote statistical significance p<0.001 in one-sample t-tests.

DL (Radar): We take inspiration from [38, 49] to implement a standard 1-D Convolutional neural network (CNN) architecture for respiratory estimation from the beamformed samples captured by the radar. While the authors of the original work employed the CNN for HR estimation, we place our focus on calculating RR. Prior to processing the radar data, we beamform the radar signal, and take the Fourier transform along the fast-time axis, similar to [38, 49]

Fusion Network: We perform fusion in two steps. First, we train a Siamese Contrast-Phys+ model from scratch (without pretraining) to fuse the two thermal cameras. The output waveform of the resulting camera-fusion model is concatenated with the output of the RF-Net and the resulting 2-channel resp waveform , i.e., the 2-channel time-series data is processed by a 1D-CNN to yield the final resp waveform. All respiration models (thermal, radar, and waveform fusion) are trained from scratch with modality-specific frequency losses; camera-fusion uses mid-level fusion, while camera+radar employs late waveform fusion.

4.3 EO Signal, BR and BRV

From the preprocessed RGB (left stereo) video frames, we extract the EO signal with the help of Google's *mediapipe* [25] library. *Mediapipe* is used to extract the eyelid landmarks per frame. To normalize the EO signal, we divide the vertical opening of each eye by its horizontal spread by detecting four landmarks per eye. Lastly, we average the EO signal across both eyes, and invert this normalized EO signal and extract BR and BRV in the same way as the extraction of HR and HRV from PPG waveforms is done.

5 Estimating Cognitive Load

5.1 PPG and Respiratory Feature Extraction

A comprehensive suite of physiological features is extracted from contact-based PPG and respiratory signals to model cardiovascular and respiratory dynamics associated with cognitive load. PPG signals undergo Butterworth bandpass filtering $(0.8 - 3.0\ Hz)$ and subsequent peak detection to yield inter-beat intervals (IBIs). From these IBIs, diverse HRV metrics are computed. These include time-domain indices quantifying beat-to-beat fluctuations (e.g., RMSSD, SDNN), frequency-domain metrics reflecting autonomic nervous system balance (e.g., LF/HF ratio), geometric HRV assessments from Poincaré plots, and descriptors of pulse wave morphology. For respiratory signals, after similar bandpass filtering $(0.05 - 1.0\ Hz)$, extracted features include the dominant frequency, the distribution of signal power across the respiratory frequency band, and spectral entropy to quantify breathing pattern regularity. Linear interpolation is employed for occasional missing data.

5.2 ML Framework for Cognitive Load Classification

For cognitive load classification, we benchmarked both traditional ML models and DL architectures. The ML suite, including Random Forest (RF) and Gradient Boosting (GB), Support Vector Machine (SVM), Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-nearest neighbor (KNN), Decision Trees (DT), and multilayer perceptrons (MLP), operated on the standardized engineered physiological features previously described. The DL approaches, including CNN, LSTM, and ResNet, processed and scaled raw physiological signal segments directly. These evaluations serve as initial benchmarks for the CogPhys dataset, providing baseline performance metrics across various models and feature types. This is intended to facilitate standardized comparison and guide future algorithmic development by users of this new multimodal resource. We systematically assessed performance across seven distinct physiological signal combinations, ranging from unimodal inputs (Contact PPG Only, rPPG Only, Blink Markers Only) through contact-remote mixtures to a full multimodal remote setup (rPPG + Remote Respiratory + Blink Markers). This experimental design allowed for a comparison of contact versus remote sensing.

6 Results

6.1 Training and Evaluation Configuration

All recordings are 2 $mins$ long. Data modalities required for cardiac waveform estimation are resampled to 30 Hz, while all respiratory-related recordings are resampled to 15 Hz. Further, all camera frames are downsized to 128×128. Single-channel inputs are replicated to thrice. For the radar, we consider the first 10 range bins. We use hierarchical windows to balance detail & robustness:

1. Training: Waveform regression models are trained on 300-sample length clips. This corresponds to 10 $secs$ for PPG at 30 Hz, and 20 $secs$ for respiratory at 15 Hz.
2. Vital sign estimation: HR computed on 30 sec windows, RR on 40 sec windows. This is carried out by concatenating the waveforms from the regression models.
3. Cognitive load classification: For each recording, all the estimated waveforms from the regression models are concatenated to form the entire 2 min window. The features are then extracted from the entire 2 min waveform.

Ground Truth Calculation: Contact sensors (pulse oximeter, chest band) provide $2 - min$ waveforms. We apply frequency peak detection on non-overlapping windows, yielding 4 HR and 3 RR values per recording. Each 2 min recording receives a single cognitive load label from NASA-TLX surveys

6.2 Remote Vitals Estimation

HR and HRV Estimation: Table 2 presents the quantitative performance of the various rPPG algorithms. Due to the SNR difference between the visible and NIR wavelengths, RGB methods outperform the NIR across all methods. However, this gap in performance can be bridged by pretraining the NIR model using the RGB model weights, with Contrast-Phys+ demonstrating improvements $\sim 0.8\ BPM$. We surmise that low SNR of the NIR signals necessitates a "head start" via pretraining. However, this mainly improves frequency preservation, with HRV observing marginal gains. Following this logic, we pre-trained the fusion network using the NIR models' weights. The resulting fusion algorithm outperformed all unimodal techniques, achieving average errors $< 3\ BPM$ on HR estimation, and also HRV estimation, with an error of $< 32\ ms$. An analysis of the standard errors indicates that the performance of the algorithms is within a 10% spread of the MAE and IBI for most models, especially the top performing models. An in-depth analysis with cross-fold validation, significance testing and skin-tone bias quantification is included in the supplement. The supplement also include metrics for clinical significance [11] and waveform characterization.

RR Estimation: The typical range of RR when seated varies between 10–30 RPM, resulting in smaller performance absolute gains compared to rPPG. From Table 3, the thermal camera placed below (TB) (directly viewing nostrils), the cockpit performs better than the thermal camera placed above (TA) across a majority of models (measuring diffused temperature changes near nostrils). The radar achieves a lower MAE compared to both unimodal camera-based approaches and camera-fusion. The waveform fusion, however, emerges as the best configuration over the unimodal and

Table 2: **Performance of rPPG algorithms on the *CogPhys* dataset.** We report the mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and the Pearson correlation (r) for HR estimation. We also report the performance of HRV estimation. The standard error spread for each metric has also been tabulated. **Clinical and waveform metrics have been tabulated in the supplement.** Post-processing steps were used to clean the waveforms before error calculation. The best and second-best numbers are shown in **bold** and underline respectively.

		HR Metrics (BPM)				HRV Metric
	Method	MAE ↓	RMSE ↓	MAPE ↓	r ↑	IBI (ms) ↓
NIR	**PhysNet** [45]	12.61 ± 0.61	15.68 ± 4.57	15.44 ± 0.68	0.05 ± 0.07	332.93 ± 11.74
	RythmFormer [50]	7.47 ± 0.48	10.53 ± 3.48	9.08 ± 0.57	0.54 ± 0.05	71.50 ± 3.89
	PhysFormer [47]	10.82 ± 0.56	13.81 ± 4.27	13.49 ± 0.67	0.06 ± 0.07	104.10 ± 5.13
	FactorizePhys [18]	10.41 ± 0.82	16.37 ± 5.57	12.22 ± 0.92	0.35 ± 0.06	80.33 ± 4.68
	PhysMamba [26]	5.47 ± 0.52	9.64 ± 3.93	6.69 ± 0.64	0.63 ± 0.05	65.00 ± 3.98
	Contrast-Phys+ [36]	5.55 ± 0.49	9.32 ± 3.55	6.79 ± 0.58	0.65 ± 0.05	59.87 ± 3.86
	Pretrained Contrast-Phys+	4.77 ± 0.45	8.44 ± 3.49	5.84 ± 0.54	0.72 ± 0.05	56.51 ± 3.54
RGB	**PhysNet** [45]	12.37 ± 0.60	15.45 ± 4.53	15.38 ± 0.72	0.12 ± 0.06	363.35 ± 12.75
	RythmFormer [50]	7.54 ± 0.47	10.49 ± 3.44	9.31 ± 0.56	0.51 ± 0.06	72.00 ± 3.72
	PhysFormer [47]	6.36 ± 0.43	9.15 ± 3.18	7.91 ± 0.52	0.63 ± 0.05	56.77 ± 3.37
	FactorizePhys [18]	5.87 ± 0.70	12.24 ± 5.56	6.71 ± 0.73	0.52 ± 0.06	47.00 ± 3.41
	PhysMamba [26]	4.08 ± 0.43	7.79 ± 3.57	4.82 ± 0.46	0.77 ± 0.04	$\underline{35.18} \pm 2.27$
	Contrast-Phys+ [36]	$\underline{3.75} \pm 0.38$	$\underline{6.94} \pm 3.23$	$\underline{4.41} \pm 0.41$	$\underline{0.83} \pm 0.04$	37.90 ± 2.93
	Fusion	**2.94** ± 0.27	**5.03** ± 2.34	**3.60** ± 0.30	**0.9** ± 0.03	**31.93** ± 2.10

Table 3: **Performance of resp. rate (RR) estimation algorithms on the *CogPhys* dataset.** We report the MAE, MAPE, RMSE, and the Pearson correlation. The standard error spread for each metric has also been tabulated. **Waveform metrics have been tabulated in the supplement.** Post-processing steps were used to clean the waveforms before error calculation. The best and second-best performing numbers are shown in **bold** and underline, respectively.

		RR Metrics (RPM)			
	Method	MAE ↓	RMSE ↓	MAPE ↓	r ↑
Above	**PhysNet** [45]	3.26 ± 0.20	4.22 ± 1.44	19.94 ± 1.41	-0.04 ± 0.08
	RythmFormer [50]	2.46 ± 0.16	3.21 ± 1.10	15.33 ± 1.11	0.08 ± 0.08
	PhysFormer [47]	3.67 ± 0.22	4.65 ± 1.56	20.66 ± 1.12	-0.11 ± 0.08
	FactorizePhys []	3.27 ± 0.23	4.50 ± 1.56	20.35 ± 1.69	-0.07 ± 0.08
	PhysMamba [26]	2.85 ± 0.21	4.00 ± 1.51	17.23 ± 1.45	-0.01 ± 0.08
	Contrast-Phys+ [36]	2.46 ± 0.18	3.42 ± 1.23	14.74 ± 1.14	0.15 ± 0.08
Below	**PhysNet** [45]	2.49 ± 0.20	3.63 ± 1.32	15.01 ± 1.31	0.13 ± 0.08
	RythmFormer [50]	2.58 ± 0.20	3.67 ± 1.32	14.91 ± 1.19	0.09 ± 0.08
	PhysFormer [47]	3.44 ± 0.23	4.59 ± 1.50	19.39 ± 1.25	-0.04 ± 0.08
	FactorizePhys [18]	2.83 ± 0.24	4.22 ± 1.65	17.32 ± 1.56	0.10 ± 0.08
	PhysMamba [26]	2.40 ± 0.20	3.54 ± 1.30	14.16 ± 1.20	0.08 ± 0.08
	Contrast-Phys+ [36]	$\underline{2.27} \pm 0.21$	3.58 ± 1.57	$\underline{13.58} \pm 1.34$	**0.31** ± 0.07
RF	**RF-Net**	2.32 ± 0.17	$\underline{3.19} \pm 1.14$	14.12 ± 1.16	0.16 ± 0.08
	Cameras Fusion	2.41 ± 0.19	3.51 ± 1.25	14.07 ± 1.17	0.07 ± 0.08
	Waveform Fusion	**2.25** ± 0.17	**3.15** ± 1.15	**13.36** ± 1.02	$\underline{0.23} \pm 0.07$

camera-fusion models, albeit only marginally. Here, we note that RR calculation can inherently be erroneous. Respiratory waveforms deviate from their periodicity when in the presence of motion or speech, and the GT also lacks a dominant frequency peak. An in-depth analysis with cross-fold validation and significance testing is included in the supplement along with waveform metrics.

6.3 Cognitive Load Classification

Evaluation of ML for cognitive load classification revealed that multimodal signal integration is highly effective. As detailed in Table 4, ML models using engineered physiological features consistently outperformed DL approaches. The GB classifier achieved the highest overall accuracy of 86.49% (F1: 0.878) with a full multimodal set comprising rPPG, Remote Respiratory, and Blink Markers. This significantly surpassed unimodal results (e.g., rPPG alone at an accuracy of 69.23% with RF). The inclusion of blink markers notably enhanced performance, underscoring the value of ocular dynamics.

Table 4: **Performance comparison of ML and DL models across various unimodal and multimodal physiological signal combinations for cognitive load classification.** In each cell we report the Accuracy(F1 Score). **Bold** values indicate the highest accuracy for each ML model; underlined values indicate second-highest. The overall best model is marked in red and the second best in blue

Model	Contact PPG	Remote PPG	Blink Markers	Contact PPG + Contact Resp	Contact PPG + Contact Resp + Blink Markers	rPPG + Contact Resp	rPPG + Remote Resp	rPPG + Remote Resp + Blink Markers
ML Models								
RF	0.70(0.73)	0.56(0.59)	0.65(0.70)	0.65(0.68)	0.73(0.76)	0.56(0.59)	0.65(0.68)	**0.78**(0.80)
GB	0.58(0.56)	0.69(0.73)	0.63(0.67)	0.60(0.64)	0.78(0.80)	0.69(0.73)	0.57(0.58)	**0.86**(0.88)
SVM	0.60(0.62)	0.51(0.56)	0.58(0.65)	0.65(0.70)	**0.80**(0.83)	0.51(0.56)	0.57(0.56)	0.76(0.80)
LR	0.63(0.63)	0.64(0.70)	0.58(0.65)	0.58(0.56)	0.80(0.83)	0.64(0.70)	0.54(0.51)	**0.81**(0.84)
LDA	0.50(0.44)	0.62(0.65)	0.50(0.57)	0.48(0.49)	**0.80**(0.83)	0.62(0.65)	0.57(0.56)	0.70(0.73)
KNN	0.45(0.50)	0.59(0.58)	0.55(0.53)	0.50(0.60)	0.68(0.75)	0.59(0.58)	0.51(0.57)	**0.70**(0.78)
DT	0.53(0.54)	0.51(0.54)	0.48(0.57)	0.63(0.68)	0.58(0.60)	0.51(0.54)	0.62(0.61)	**0.76**(0.77)
MLP	0.55(0.55)	0.62(0.59)	0.60(0.67)	0.60(0.60)	**0.88**(0.88)	0.62(0.59)	0.54(0.48)	0.76(0.80)
DL Models								
1D CNN	0.65(0.67)	0.49(0.00)	**0.68**(0.70)	0.50(0.67)	0.50(0.67)	0.56(0.26)	0.49(0.10)	0.54(0.70)
LSTM	0.50(0.50)	0.59(0.64)	0.58(0.62)	0.50(0.67)	0.50(0.67)	0.54(0.50)	0.51(0.55)	**0.62**(0.74)
ResNet1D	0.63(0.59)	0.59(0.56)	**0.68**(0.71)	0.50(0.67)	0.50(0.67)	0.54(0.50)	0.49(0.63)	0.62(0.68)

To validate the effectiveness of remote sensing, we evaluated the Contact PPG + Contact Resp + Blink Markers combination as a practical upper bound. This combination achieved 87.5% accuracy (F1: 0.884) with MLP and 80.0% accuracy (F1: 0.826) with SVM. The similar test set performance between this contact-based multimodal approach (87.5%) and our best remote-based multimodal approach (86.49%) demonstrates that integrated remote sensing can achieve near-equivalent results to traditional contact methods, supporting the feasibility of non-intrusive cognitive monitoring systems.

On the other hand, DL architectures like 1D CNNs performed best with only Blink Markers (68.00% accuracy) and were less effective with multimodal inputs compared to traditional ML. This suggests a current limitation in DL models for integrating heterogeneous physiological signals on this dataset.

A significant finding was that the combination of multiple remote sensing modalities ultimately delivered the best performance, even surpassing individual contact-based sensor setups. This highlights the potential of integrated remote sensing. Ensemble ML methods, particularly GB, proved most robust. Our participant-based test split ensures these findings are generalizable, validating real-world deployment potential without individual calibration.

7 Conclusion and Limitations

We introduced CogPhys, an open-source, multimodal dataset featuring diverse remote (RGB, NIR, thermal, radar) and contact sensors for advancing remote cognitive load assessment. Our benchmarks demonstrate successful remote physiological signal estimation and establish strong baselines for cognitive load classification, achieving up to 86.49% accuracy with multimodal remote signals and highlighting the utility of blink-related features. CogPhys provides a valuable public resource to develop and evaluate new algorithms for less intrusive cognitive state monitoring.

Key limitations include data collection in a controlled laboratory setting, which may not fully reflect real-world complexities. While our experimental design focused on inducing cognitive load through tasks validated in prior work [10, 12], potential confounding with stress responses remains an important consideration for future investigation. The absence of direct brain activity sensors (EEG/fNIRS) represents another limitation, though our validation against self-reported NASA-TLX scores provides a practical ground truth. Furthermore, cognitive load was defined via specific tasks and binarized using personalized NASA-TLX thresholds, offering opportunities for future work on broader task types and more granular load measures. The current dataset's participant demographics and sensor specifics also define a scope for future studies on broader generalizability. These limitations point towards clear directions for extending this research.

References

[1] Mark B. Abelson. It's time to think about the blink. `https://www.reviewofophthalmology.com/article/its-time-to-think-about-the-blink`, 2011. Accessed: 2025-05-15.

[2] Christoph Hoog Antink, Hanno Gao, Christoph Brüser, and Steffen Leonhardt. Beat-to-beat heart rate estimation fusing multimodal video and sensor data. *Biomedical optics express*, 6(8):2895–2907, 2015.

[3] Karina Rollandovna Arutyunova, Anastasiia Vladimirovna Bakhchina, Daniil Igorevich Konovalov, Mane Margaryan, Andrei Viktorovich Filimonov, and Ivan Sergeevich Shishalov. Heart rate dynamics for cognitive load estimation in a driving simulation task. *Scientific Reports*, 14(1):31656, 2024.

[4] Guha Balakrishnan, Fredo Durand, and John Guttag. Detecting pulse from head motions in video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3430–3437, 2013.

[5] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern recognition letters*, 124:82–90, 2019.

[6] Pradyumna Chari, Anirudh Bindiganavale Harish, Adnan Armouti, Alexander Vilesov, Sanjit Sarda, Laleh Jalilian, and Achuta Kadambi. Implicit neural models to extract heart rate from video. In *European conference on computer vision*, pages 157–175. Springer, 2024.

[7] Pradyumna Chari, Krish Kabra, Doruk Karinca, Soumyarup Lahiri, Diplav Srivastava, Kimaya Kulkarni, Tianyuan Chen, Maxime Cannesson, Laleh Jalilian, and Achuta Kadambi. Diverse r-ppg: Camera-based heart rate estimation for diverse subject skin-tones and scenes. *arXiv preprint arXiv:2010.12769*, 2020.

[8] Youngjun Cho, Simon J Julier, Nicolai Marquardt, and Nadia Bianchi-Berthouze. Robust tracking of respiratory rate in high-dynamic range scenes using mobile thermal imaging. *Biomedical optics express*, 8(10):4480–4503, 2017.

[9] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE transactions on biomedical engineering*, 60(10):2878–2886, 2013.

[10] Cary Deck, Salar Jahedi, and Roman Sheremeta. On the consistency of cognitive load. *European Economic Review*, 134:103695, 2021.

[11] Association for the Advancement of Medical Instrumentation et al. Cardiac monitors, heart rate meters, and alarms. *American National Standard (ANSI/AAMI EC13: 2002) Arlington, VA*, pages 1–87, 2002.

[12] Holger Gerhardt, Guido P Biele, Hauke R Heekeren, and Harald Uhlig. Cognitive load increases risk aversion. Technical report, SFB 649 Discussion Paper, 2016.

[13] Martin Gjoreski, Mitja Luštrek, and Veljko Pejović. My watch says i'm busy: inferring cognitive load with low-cost wearables. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pages 1234–1240, 2018.

[14] Mariel Grassmann, Elke Vlemincx, Andreas Von Leupoldt, Justin M Mittelstädt, and Omer Van den Bergh. Respiratory changes in response to cognitive load: A systematic review. *Neural plasticity*, 2016(1):8146809, 2016.

[15] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[16] Morris K Holland and Gerald Tarlow. Blinking and mental load. *Psychological Reports*, 31(1):119–127, 1972.

[17] Wonse Jo, Ruiqi Wang, Go-Eum Cha, Su Sun, Revanth Krishna Senthilkumaran, Daniel Foti, and Byung-Cheol Min. Mocas: A multimodal dataset for objective cognitive workload assessment on simultaneous tasks. *IEEE Transactions on Affective Computing*, 2024.

[18] Jitesh Joshi, Sos S Agaian, and Youngjun Cho. Factorizephys: Matrix factorization for multidimensional attention in remote physiological sensing. *arXiv preprint arXiv:2411.01542*, 2024.

[19] Jitesh Joshi and Youngjun Cho. Ibvp dataset: Rgb-thermal rppg dataset with high resolution signal quality labels. *Electronics*, 13(7):1334, 2024.

[20] Emmanouil Ktistakis, Vasileios Skaramagkas, Dimitris Manousos, Nikolaos S Tachos, Evanthia Tripoliti, Dimitrios I Fotiadis, and Manolis Tsiknakis. Colet: A dataset for cognitive workload estimation based on eye-tracking. *Computer Methods and Programs in Biomedicine*, 224:106989, 2022.

[21] Mayank Kumar, Ashok Veeraraghavan, and Ashutosh Sabharwal. Distanceppg: Robust non-contact vital signs monitoring using a camera. *Biomedical optics express*, 6(5):1565–1588, 2015.

[22] Xiaotian Li, Zheng Zhang, Xiang Zhang, Taoyue Wang, Zhihua Li, Huiyuan Yang, Umur Ciftci, Qiang Ji, Jeffrey Cohn, and Lijun Yin. Disagreement matters: Exploring internal diversification for redundant attention in generic facial action analysis. *IEEE Transactions on Affective Computing*, 15(2):620–631, 2023.

[23] Xin Liu, Girish Narayanswamy, Akshay Paruchuri, Xiaoyu Zhang, Jiankai Tang, Yuzhe Zhang, Roni Sengupta, Shwetak Patel, Yuntao Wang, and Daniel McDuff. rppg-toolbox: Deep remote ppg toolbox. *Advances in Neural Information Processing Systems*, 36:68485–68510, 2023.

[24] Yiming Liu, Binjie Qin, Rong Li, Xintong Li, Anqi Huang, Haifeng Liu, Yisong Lv, and Min Liu. Motion-robust multimodal heart rate estimation using bcg fused remote-ppg with deep facial roi tracker and pose constrained kalman filter. *IEEE Transactions on Instrumentation and Measurement*, 70:1–15, 2021.

[25] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019.

[26] Chaoqi Luo, Yiping Xie, and Zitong Yu. Physmamba: Efficient remote physiological measurement with slowfast temporal difference mamba. In *Chinese Conference on Biometric Recognition*, pages 248–259. Springer, 2024.

[27] Alfonso Magliacano, Salvatore Fiorenza, Anna Estraneo, and Luigi Trojano. Eye blink rate increases as a function of cognitive load during an auditory oddball paradigm. *Neuroscience Letters*, 736:135293, 2020.

[28] Daniel McDuff, Miah Wander, Xin Liu, Brian Hill, Javier Hernandez, Jonathan Lester, and Tadas Baltrusaitis. Scamps: Synthetics for camera measurement of physiological signals. *Advances in Neural Information Processing Systems*, 35:3744–3757, 2022.

[29] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14*, pages 562–576. Springer, 2019.

[30] Ewa Nowara, Tim K Marks, Hassan Mansour, and Ashok Veeraraghavan. Sparseppg: Towards driver monitoring using camera-based vital signs estimation in near-infrared. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1272–1281, 2018.

[31] Ewa M Nowara, Tim K Marks, Hassan Mansour, and Ashok Veeraraghavan. Near-infrared imaging photoplethysmography during driving. *IEEE transactions on intelligent transportation systems*, 23(4):3589–3600, 2020.

[32] Marcus Nyström, Richard Andersson, Diederick C Niehorster, Roy S Hessels, and Ignace TC Hooge. What is a blink? classifying and characterizing blinks in eye openness signals. *Behavior research methods*, 56(4):3280–3299, 2024.

[33] Amruta Pai, Ashok Veeraraghavan, and Ashutosh Sabharwal. Camerahrv: robust measurement of heart rate variability using a camera. In *Optical Diagnostics and Sensing XVIII: Toward Point-of-Care Diagnostics*, volume 10501, pages 160–168. SPIE, 2018.

[34] Soroosh Solhjoo, Mark C Haigney, Elexis McBee, Jeroen JG van Merrienboer, Lambert Schuwirth, Anthony R Artino Jr, Alexis Battista, Temple A Ratcliffe, Howard D Lee, and Steven J Durning. Heart rate and heart rate variability correlate with clinical reasoning performance and self-reported measures of cognitive load. *Scientific reports*, 9(1):14668, 2019.

[35] Zhaodong Sun and Xiaobai Li. Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast. In *European Conference on Computer Vision*, pages 492–510. Springer, 2022.

[36] Zhaodong Sun and Xiaobai Li. Contrast-phys+: Unsupervised and weakly-supervised video-based remote physiological measurement via spatiotemporal contrast. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[37] Jiankai Tang, Kequan Chen, Yuntao Wang, Yuanchun Shi, Shwetak Patel, Daniel McDuff, and Xin Liu. Mmpd: Multi-domain mobile video physiology dataset. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–5. IEEE, 2023.

[38] Alexander Vilesov, Pradyumna Chari, Adnan Armouti, Anirudh Bindiganavale Harish, Kimaya Kulkarni, Ananya Deoghare, Laleh Jalilian, and Achuta Kadambi. Blending camera and 77 ghz radar sensing for equitable, robust plethysmography. *ACM Trans. Graph.*, 41(4):36–1, 2022.

[39] Wenjin Wang, Albertus C Den Brinker, Sander Stuijk, and Gerard De Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2016.

[40] Zhen Wang, Yunhao Ba, Pradyumna Chari, Oyku Deniz Bozkurt, Gianna Brown, Parth Patwa, Niranjan Vaddi, Laleh Jalilian, and Achuta Kadambi. Synthetic generation of face videos with plethysmograph physiology. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20587–20596, 2022.

[41] Christopher D Wickens. Multiple resources and mental workload. *Human factors*, 50(3):449–455, 2008.

[42] Zheng Wu, Yiping Xie, Bo Zhao, Jiguang He, Fei Luo, Ning Deng, and Zitong Yu. Cardiac-mamba: A multimodal rgb-rf fusion framework with state space models for remote physiological measurement. *arXiv preprint arXiv:2502.13624*, 2025.

[43] Ronglong Xiong, Fanmeng Kong, Xuehong Yang, Guangyuan Liu, and Wanhui Wen. Pattern recognition of cognitive load using eeg and ecg signals. *Sensors*, 20(18):5122, 2020.

[44] Mark S Young, Karel A Brookhuis, Christopher D Wickens, and Peter A Hancock. State of science: mental workload in ergonomics. *Ergonomics*, 58(1):1–17, 2015.

[45] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *arXiv preprint arXiv:1905.02419*, 2019.

[46] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Yawen Cui, Jiehua Zhang, Philip Torr, and Guoying Zhao. Physformer++: Facial video-based physiological measurement with slowfast temporal difference transformer. *International Journal of Computer Vision*, 131(6):1307–1330, 2023.

[47] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip HS Torr, and Guoying Zhao. Physformer: Facial video-based physiological measurement with temporal difference transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4186–4196, 2022.

[48] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3438–3446, 2016.

[49] Tianyue Zheng, Zhe Chen, Shujie Zhang, Chao Cai, and Jun Luo. More-fi: Motion-robust and fine-grained respiration monitoring via deep-learning uwb radar. In *Proceedings of the 19th ACM conference on embedded networked sensor systems*, pages 111–124, 2021.

[50] Bochao Zou, Zizheng Guo, Jiansheng Chen, Junbao Zhuo, Weiran Huang, and Huimin Ma. Rhythmformer: Extracting patterned rppg signals based on periodic sparse attention. *Pattern Recognition*, 164:111511, 2025.

[51] Bochao Zou, Zizheng Guo, Xiaocheng Hu, and Huimin Ma. Rhythmmamba: Fast remote physiological measurement with arbitrary length videos. *arXiv preprint arXiv:2404.06483*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper proposes a new dataset consisting of multi-modal recordings of seated participants doing tasks of varying cognitive loads. Remote vital signs and other physiological signals are estimated from such recordings, which are then further used for cognitive load estimation. This aligns with the claims made in the abstract and introduction.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes] .

   Justification: We discuss this in a separate section in the supplement.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper's contributions do not involve any theoretical results, which require proof and assumptions.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide links to the dataset and code used in our experiments. Dedicated README files have been included in both to further explain the order and commands to run the code. Further, we will be releasing our trained models, from which the results can be reproduced.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in the supplemental material?

   Answer: [Yes]

   Justification: As per the discussion from the rebuttal phase, the dataset will be gated via a Data User Agreement (DUA) due to privacy-related reasons. Upon signing the document, researchers will be provided with a link to our dataset. This is in line with the suggestion from the Reviewers. We will be the codebase for vital signs estimation and cognitive load estimation in GitHub: https://github.com/AnirudhBHarish/CogPhys.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We include these details in both the supplement and the accompanying code. We also include a pickle file with the data split we have used. README files also include instructions for running the code bases.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Our supplement includes statistical significance testing results (ANOVA and Tukey HSD tests) for heart rate (HR), respiratory rate (RR), and cognitive load classification.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide these details in the supplemental material.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have followed the guidelines as per the NeurIPS Code of Ethics. All data was collected under IRB approval, with consent obtained from all participants.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the applications of our work both in the main paper and the supplement. The paper highlights positive societal impacts, such as the development of more user-friendly and less intrusive physiological measurement techniques by replacing contact-based sensors.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for the responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We have collected data under IRB approval, and with signed consent from all participants. Per discussion with the reviewers, we will be gating the dataset with a DUA for privacy-related reasons.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited, and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited and acknowledged the creators or original owners of the used assets.

13. **New assets**

    Question: Are new assets introduced in the paper well documented, and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We explain all technical and logistical details relating to the dataset in the main paper and Supplement. Higher-level details, such as sensor choice and acquisition protocols, are discussed in the main paper, while the lower-level technical details, such as pre- and post-processing algorithms, along with data specification, have been elaborated on in the Supplement. The dataset and codebase links accompanying the paper will also contain this information.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: We provide these details in Section 3.3 in the main paper and the supplemental material.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: We provide these details in Section 3.3 and the supplemental material.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, a declaration is not required.

Answer: [NA]

Justification: