Supplement - 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

1 Dataset Design

1.1 Dataset Specifications

Following the protocols elaborated in the Section 3.3, we collected a large-scale dataset, dubbed the *CogPhys* dataset. We recruited 37 participants for our study, and the dataset was collected in compliance with an Institutional Review Board (IRB). Our protocol was approved by the Rice University Institutional Review Board under the study: IRB-FY2025-59.

Across the dataset, we have six 2 min recordings from each participant, barring two participants, from whom we have five recordings. This yielded to a total of 220 recordings, which totals to 440 $mins$ of video and radar recordings. Furthermore, the raw recordings for each task were $\sim 42.5\,GBs$, summing up to $\sim\ 255\,GBs$ per participant. Due to the size of the dataset, we resize, crop, and compress the data (losslessly) to approximately $2 - 2.5GB$ per task. We elaborate on the preprocessing and compression steps in the next section. **Due to the size of the raw dataset** $(> 9\ TBs)$**, we will be releasing the preprocessed and compressed dataset** $\sim 600\ GBs$. In addition to the contact and remote sensor recordings, we also collect demographic data from the participants. These include the use of glasses/contacts, use of cosmetics, age, gender, self-reported Fitzpatrick skin tone values, and height.

For all our experiments, we partition the dataset into 3 sets - train, test, and validation sets. We split the dataset according to participants (a split corresponding to 25/10/2 for train/test/val), ensuring there is no overlap between the train, test, and validation sets. All metrics and results reported in this paper are derived from the test set. We have included a pickle file with the train, test, and validation split in the codebase accompanying the paper.

1.2 Demographic Distribution and Bias Effects

A summary of the dataset demographics and specifications has been listed in Table 1. Given our participant size of 37 and distribution statistics, our dataset has limited representation across all demographic groups. As such, algorithms trained on this dataset would reflect the bias present in our demographic distribution. Furthermore, rPPG is known to be biased against demographics with

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

Table 1: Summary of Dataset Statistics

Statistic	**Value/Count**
Total Participants	37
Gender (Male)	23
Gender (Female)	14
Mean Age	24.09
Median Age	25
Age Range	$18 - 31$
Lens/Contacts (Y)	19
Lens/Contacts (C)	7
Lens/Contacts (N)	11
Uses Cosmetics	6
Does Not Use Cosmetics	31
Skin Tone (fp2)	10
Skin Tone (fp3)	11
Skin Tone (fp4)	10
Skin Tone (fp5)	5
Missing Skin Tone	1
Mean Height	$5'6\frac{2}{3}''$
Median Height	$5'6''$
Height Range	$5'0'' - 6'3''$

darker skin tones [44, 34] on an imaging level. We present our analysis of the same in Section 6.6 of this supplement. De-biasing is an important topic within remote physiological sensing. While the main goal our paper was to explore remote cognitive load estimation, future works can dive deeper into the demographic effects and algorithms to debias the same.

1.3 Data Preprocessing

The radar, NIR, and RGB cameras are set to a sampling rate of $30\ Hz$, while the thermal cameras and pulse oximeter are operated at their default sampling rate of $60\ Hz$. The respiratory band, on the other hand, samples the respiratory signals at $18\ Hz$ and the electrocardiography (ECG) signal at $100\ Hz$. The accumulated data packets are transmitted to the computer at $1\ Hz$. We store the raw pulse oximeter and ECG recordings (when available). The respiratory signal is upsampled to $30\ Hz$ prior to consolidation.

The data from the NIR and thermal cameras is captured with 16-bit integer depth. The raw in-phase quadrature (IQ) data from the radar is captured with all 3 transmitters and 4 receivers active, and with each data frame containing 8 chirps. The only pre-processing step performed for the radar is to rearrange the captured raw data into a radar matrix. Mathematical details for remote photoplethysmography (rPPG) signal formation and radar matrix acquisition can be found in [45] and [44], respectively. A detailed list of camera and radar parameters and data shapes is provided in the supplement.

In contrast, the raw video recordings are cropped to the face and compressed. We use Google's *mediapipe* [26] library to crop the faces from the camera-based modalities. We find that *mediapipe's* facial landmarks detection function can be used on all 3 camera modalities - RGB, NIR, and thermal (placed above the cockpit). Due to the complexity of the tasks and the nature of the thermal videos, we only estimate the landmarks for frames where the landmark detector can estimate the keypoints. We pool the estimated landmarks to draw a single bounding box around the face, such that the participant's face is within this box across all valid frames. Due to the extreme viewing angle of the thermal camera placed below the cockpit, we implement a hand-crafted adaptive thresholding algorithm to iteratively threshold the video. The final thresholded output from the iterative process is used to draw the facial bounding box. The cropped frames from all 4 cameras are then resized to a constant size: 256×256.

Further down, these frames can be downsized for the processing algorithms, as in our case where the frames were downsampled to $128x128$ to train the models. All sensing processes were started

concurrently via multiprocessing, and per-sample receive timestamps were logged on the acquisition laptop for alignment post hoc. Further details are provided in the dataset.

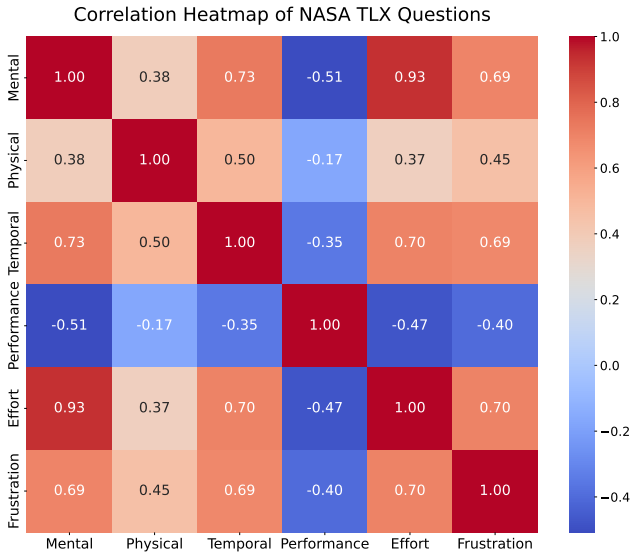


Figure 1: Correlation heatmap between the six NASA-TLX dimensions, showing strong positive correlations between Mental Demand, Temporal Demand and frustration ($r \approx 0.7$), while Performance shows negative correlations with all other dimensions, suggesting that higher perceived workload corresponds to lower self-rated performance.

2 Extended Literature Review

This section provides an extended review of the literature that complements the concise related work presented in the main manuscript. We cover the evolution of remote physiological sensing technologies, multimodal fusion approaches, and cognitive state estimation methodologies that form the foundation of our work.

2.1 Remote Unimodal Vital Sensing

Heart rate (HR), heart rate variability (HRV), and respiratory rate (RR) sensing with cameras and radars have gained great popularity [46, 35, 7]. This has led to an uptick in large-scale datasets for training and testing the various algorithms for rPPG and remote respiration research. These datasets span a range of modalities and conditions, including lab environments (e.g., UBFC [6], MR-NIRP [32]), driving [33], and synthetic settings (e.g., SCAMPS [29], UCLA-rPPG [46]). These datasets have led to the development of various model-based approaches entailing spatial signal to noise ratio (SNR) maps [23, 7], sparse spectral methods for NIR imaging [33, 32], light transport-based methods [10, 45], and motion-based methods [4]. Unlike model-based approaches, deep learning (DL) adopts a data-driven approach to estimate the rPPG signal. Advanced architecture such as transforms [52, 51], mamba [27, 56] and contrastive network [40, 41] have pushed the frontier of high performing rPPG algorithms. Parallelly, there have also been strides in research into respiratory signals estimation from sensors such as thermal cameras and radio-frequency (RF) sensors such as radars [8]. [54] and [44] implement novel data augmentation techniques for ultra-wideband (UWB) and frequency modulated continuous wave (FMCW) radars for RR and HR estimation, respectively.

2.2 Remote Multimodal Vital Sensing

Using a single modality for remote sensing of vitals can prove disadvantageous, leading to systems not being robust to lighting changes and being inaccurate. It could also cause inequitable outcomes in estimation errors. The usage of more than one modality has been explored in a few works. Multimodal datasets, such as MMSE [53], EquiPleth [44], iBVP [19] and MR-NIRP [32, 33] have opened the

Table 2: Existing datasets for cognitive load estimation using physiological signals. Our CogPhys dataset uniquely combines multiple remote sensing modalities with contact-based ground truth for cognitive load assessment.

Dataset	Year	Subjects	Sensing Approach	Modalities	Cognitive Tasks	Duration	Remote Sensing
Driver Workload [39]	2013	10	Contact	ECG, BTemp, SCR	Driving videos, Real driving	Variable	No
MMOD-COG [30]	2019	40	Contact	ECG, EDA, Speech	Arithmetic, Reading	Variable	No
CLAS [28]	2019	62	Contact	ECG, PPG, EDA	Math, Logic, Stroop test	Variable	No
CogLoad [15]	2020	23	Contact	HR, IBI, EDA, ST, ACC	n-back tasks, Visual cues	Variable	No
Snake [15]	2020	23	Contact	HR, IBI, EDA, ST, ACC	Snake game	Variable	No
Kalatzis et al. [20]	2021	26	Contact	ECG, RR	MATB-II	Variable	No
COLET [22]	2022	30	Hybrid	Eye-tracking, EEG, ECG	Reading, Math	45 min	Partial
MOCAS [17]	2024	40	Hybrid	Video, EEG, ECG, EDA	Multiple cognitive tasks	60 min	Partial
CL-Drive [2]	2024	21	Hybrid	EEG, ECG, EDA, Gaze	Simulated driving	Variable	Partial
CLARE [5]	2024	25	Hybrid	Video, EEG, ECG, EDA	Reading, Math, Memory	90 min	Partial
CogPhys (Ours)	**2025**	37	**Multimodal Remote**	**RGB, NIR, Thermal, Radar, PPG, ECG, Resp**	Reading, Memory, Math	12 min Each	**Yes**

doors to remote multimodal vital sensing. This has led to various advanced algorithms that have been benchmarked on both unimodal as well as multimodal data streams [41, 47]. rPPG-based methods are sensitive to face motions, and [25] attempted to reduce motion corruption by adaptively filtering the ballistocardiography (BCG) signal and rPPG signals. With a similar undertone, [3] has extended the system's capabilities to include the fusion of remote ballistocardiography (rBCG signals in addition to the non-contact rPPG and contact-based BCG.

2.3 Cognitive State Estimation

The estimation of cognitive workload, a critical aspect of human-computer interaction, increasingly leverages machine learning (ML) to interpret physiological and behavioral signals. Established contact-based methods often rely on electroencephalography (EEG) to directly measure brain activity [48, 49], or utilize wearable sensors for signals like ECG, electrodermal activity (EDA), and photoplethysmography (PPG), from which cognitive states are inferred using features such as HRV [16, 17]. Non-contact approaches primarily employ cameras for eye-tracking to analyze gaze and pupil dynamics [22], or to assess facial expressions and head movements as behavioral indicators of cognitive load [17]. While remote sensing with radar has shown promise for vital signs (as discussed in Sec. 2.1), its direct application to robust cognitive state estimation is comparatively less explored.

To overcome the limitations of individual modalities, a significant trend is the adoption of multimodal approaches, fusing data from diverse sensors [17, 5]. The development and standardized validation of such complex models, often involving sophisticated feature integration or DL techniques, have been significantly advanced by benchmark datasets like COLET and MOCAS [22, 17, 2]. However, advancing robust remote cognitive load estimation, particularly with more diverse sensor inputs and task contexts, highlights an ongoing need for comprehensive public datasets and standardized evaluation methods. Despite overall progress, achieving generalizable, real-time estimation in less constrained environments remains a key challenge due to factors like ground truth ambiguity and individual differences [21]. Our work directly addresses this by introducing a novel, extensive dataset and associated benchmarks tailored for multimodal remote cognitive load estimation, aiming to fill an important gap in current research. In our participants' responses to the NASA-TLX questionnaire, itemized scores show strong positive correlations between Mental Demand, Temporal Demand and frustration, shown in Figure 1.

2.4 Existing Datasets for Cognitive Load Estimation

The development of robust cognitive load estimation systems has been significantly advanced by the availability of benchmark datasets that provide synchronized physiological data and associated ground truth labels. Table 2 summarizes the key datasets in the literature that study cognitive load using physiological signals, highlighting the evolution from primarily contact-based sensing to emerging multimodal approaches.

Early datasets in this domain, such as Driver Workload [39] and MMOD-COG [30], focused primarily on contact-based physiological sensing using traditional wearable sensors. These datasets established the foundation for understanding the relationship between physiological responses and cognitive load but were limited by the intrusive nature of contact sensing and relatively small participant pools.

More recent datasets have begun incorporating hybrid approaches that combine contact physiological sensors with non-contact behavioral tracking. COLET [22] and MOCAS [17] represent significant advances by including eye-tracking and video analysis alongside traditional physiological measurements. However, these datasets still rely heavily on contact sensors for the primary physiological signals and have limited exploration of fully remote sensing capabilities.

Our CogPhys dataset addresses a critical gap in the literature by providing the first large-scale dataset specifically designed for multimodal remote cognitive load estimation. Unlike previous datasets that primarily rely on contact sensing or include only limited remote modalities, CogPhys incorporates five distinct remote sensing modalities (RGB cameras, NIR cameras, thermal cameras, and radar) alongside contact-based ground truth measurements. This unique combination enables systematic evaluation of remote sensing approaches while maintaining high-quality reference standards for validation.

The diversity of cognitive tasks across datasets reflects different research objectives and application domains. While some datasets focus on specific scenarios like driving [2] or gaming [15], others adopt more general cognitive load paradigms using established psychological tasks such as n-back tests [15] or mathematical problems [28]. Our dataset employs a balanced approach with reading, memorization, and arithmetic tasks that span different cognitive domains while remaining practical for real-world applications.

3 Experimental Design and Sample Retention Criteria

3.1 Design Rationale

Our task design was motivated by established cognitive load research. The protocol incorporates validated approaches from prior work [12, 14, 31, 1, 11] that induce varying cognitive load through:

- Binary grid memorization (4×4 patterns) [14, 31]
- Digit memorization (8-digit sequences) [1, 11]
- Simultaneous dual-task paradigms (memorization + arithmetic) [12]

Crucially, we allowed adequate time for both memorization and problem-solving phases to avoid inducing time pressure stress. Participants were not rushed, ensuring the tasks primarily elicited cognitive demand rather than temporal stress.

3.2 Participant Engagement Criteria

While we did not establish performance-based exclusion criteria (e.g., requiring 80% correct responses), we implemented engagement-based quality control. As documented in prior dual-task cognitive load studies [12, 37], performance on secondary tasks (our multiplication questions) naturally decreases with higher cognitive loads. Excluding low-performing participants would systematically remove those experiencing the highest cognitive load—precisely the most valuable data.

Our exclusion criteria focused on verifying genuine engagement:

- Participants who failed to attempt math problems
- Impossibly fast responses suggesting disengagement
- Clear misunderstanding of instructions (recording restarted after re-explanation)

All participants engaged genuinely with the tasks and answered validation questions correctly (e.g., simple problems like $10 \times 6 = 60$). No recordings were excluded based on engagement criteria.

4 Estimating Remote Physiological Signals from a Sensor Stack

Here we provide a brief summary of the methods used to extract the vital signs from the sensors stack. Our codebase is built on top of the rPPG-Toolbox [24] and Contrast-Phys+ [41] repositories, with one important modification. We include the *SNR Loss* from [44] as an additional loss function to

all the methods in the previously mentioned repositories. Additionally, we box-filter the estimated and ground truth waveforms prior to the Pearson loss. Filters of sizes 7and 15 were used for the rPPG and respiratory waveform regression, respectively. We find these modifications add necessary priors, drastically improving the performance on HR estimation. Additional details regarding these modifications can be found in the accompanying codebase.

4.1 Estimating rPPG and Computing HR and HRV

As a precursor to the algorithmic methods - namely Green, ICA, CHROM, and POS - we spatially average the RGB video to obtain a temporal signal of shape $T \times 3$. This is done via facial detection and segmentation using a bounding box constructed using the landmarks from the *mediapipe* [26] library. The algorithmic methods are based on exploiting the signals' redundancy across the 3 channels of the RGB videos.

Green [43]: The 1^{st} channel, i.e., the signal from the green channel, is chosen as the desired rPPG signal. This is equivalent to spatially averaging the green channel of the RGB video.

ICA [36]: Here, authors assume the rPPG signals and noise/motion signals are from independent sources. Following this assumption, they employ an independent component analysis algorithm on the 3 color channels to extract the rPPG signal.

CHROM [10]: The spatially averaged video is projected onto a chrominance space through a linear combination of the RGB channels. Following this, the chrominance signals are filtered, and the 3 color channels are combined to yield the rPPG estimate.

POS [45]: The temporal signal is projected onto a plane that is orthogonal to the skin tone. This plane is derived based on the physiological and diffuse optical properties of skin and blood flow. This essential equates to a fixed matrix projection.

The DL baselines were trained to estimate the rPPG signals through a 1-D signal regression. We supply the raw $128 \times 128 \times 3$ video frames as the input, as opposed to the default option of normalized difference frames. For the NIR video, we duplicate the channel thrice to ensure the models are identical for the RGB and NIR modalities. We add the SNR loss to all the models.

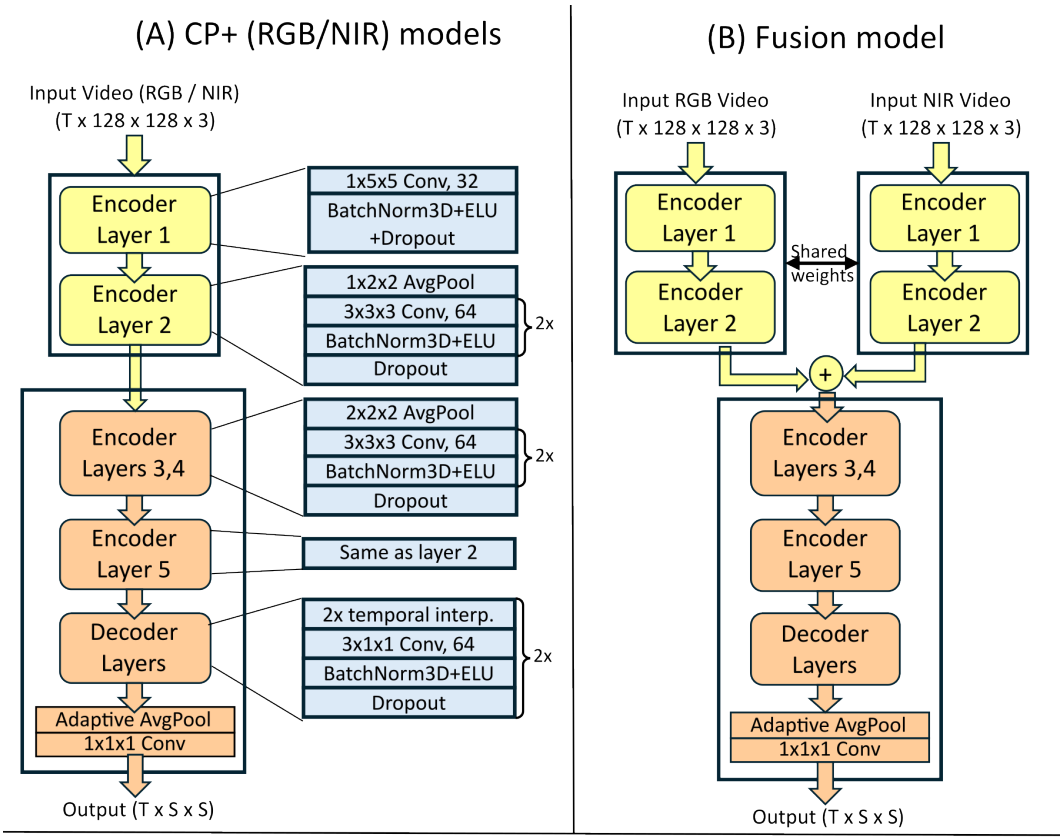
PhysNet [50]: A 3-D convolutional neural network (CNN) designed to perform a 1-D signal regression on the PPG signal. The model is trained using the negative Pearson loss function, as opposed to a traditional MSE loss.

PhysFormer [52]: A video-transformer-based architecture designed to adaptively aggregate both local and global spatio-temporal features to better represent and estimate the rPPG signal. Through a temporal-difference guided attention, they enhance the quasi-periodic features of the rPPG signal. In addition to the negative person loss function, they added a frequency domain cross-entropy loss on the frequency spectrum to classify HR values. They also added an additional label distribution term on the HR values to improve frequency domain supervision. Through curriculum learning, the frequency domain losses were emphasized towards the later epochs.

RythmFormer [55]: The authors theorize that temporal attention for periodic signals is sparse. In doing so, they introduce a periodic sparse attention mechanism to leverage this property. By filtering out redundant component pre-attention, more fine-grained features can be extracted. They use a negative Person loss and frequency domain cross-entropy loss function to train the model.

FactorizePhys [18]: A 3-D CNN architecture that incorporates matrix factorization to perform multidimensional attention from voxel embeddings. Specifically, they use NMF (nonnegative matrix factorization) to collectively compute attention across the spatial, temporal, and channel dimensions.

PhysMamba [27]: A mamba-based architecture for rPPG estimation. By leveraging the ability of state space models to efficiently model long-term dependencies, PhysMamba performs a 1-D regression to yield the rPPG signal.



Model	Training Details
CP+ (RGB model)	Trained from scratch on RGB data
CP+ (NIR model)	Initialized using CP+ (RGB); and trained on NIR data
Fusion model (ours)	Initialized using CP+ (NIR); and trained on RGB, NIR paired inputs

Figure 2: The Contrast-Phys+ model and its Siamese variant are trained for unimodal and multimodal vital sign estimation. The NIR and Fusion models are not trained from scratch. They use the pretrained weights from the RGB and NIR models, respectively. In essence, the Fusion method is the resultant model from 2 sequential pertaining stages.

Contrast-Phys+ [41]: A 3D CNN model (like PhysNet) is trained using contrastive losses on the frequency spectrum of estimated and ground truth signals[1]. The model outputs a spatio-temporal grid ($S \times S \times T$) representing multiple potential rPPG estimates. The models are trained contrastively to output similar estimates across this grid. Further, we operate the CNN with full supervision to obtain the best-performing model.

Due to the Contrast-Phys+ being the backbone of the fusion network, we make modifications to improve its overall performance. We carry this out by adding 2 loss functions to the final rPPG estimate - the spatially average of the $S \times S \times T$ grid. Following suit with the previous method, we add the SNR and negative Pearson loss functions to improve the overall performance. Additionally, we use the pre-trained weights from the RGB model as a starting point to train the NIR model as shown in Figure 2.

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

[1]In addition to an MSE loss on the normalized frequency spectrum, we also include a KL divergence loss

deep features are added and passed through the remaining layers. Due to the SNR difference between the RGB and NIR, we initialize the fusion networks using the weights of the pretrained NIR model. This gives the model a head start and improves the overall performance. Here, we notice that training the model for significant epochs can cause the RGB signals to dominate. A detailed illustration of the model architecture can be found in Figure 2.

4.2 Estimating Respiratory Signal and Computing RR

While algorithmic implementations to extract RR from thermal videos exist, they rely on facial landmarks and handcrafted features that only work well with frontal views without head rotations. Due to the complexity and chosen angles in our dataset, we rely on DL baselines to evaluate respiratory signals estimation from thermal cameras

DL (Thermal): Owing to the similarity between rPPG and remote respiratory signal sensing, we repurpose the deep-learning baselines from the previous sub-section to be compatible with respiratory signal estimation. This entails downsampling the thermal frames to 128×128 pixels and changing the frequency bands for the SNR and contrastive losses. Additionally, we downsample the input videos and ground truth signals to 15 Hz to accommodate longer window lengths. These longer windows are necessitated by the lower frequencies of breathing signals in contrast with PPG signals.

DL (Radar): Our 1-D CNN and its training routine have been adapted from the [44]. Given our task of RR estimation, we port the loss functions, while keeping the architecture the same. Additionally, we beamform the raw matrix data to strengthen the signal SNR at $0°$. We only perform the azimuth beamforming and hence make use of 2 transmitters and 4 receivers from the whole recordings. The 3rd transmitter is primarily used for elevation beamforming and not the azimuth. Taking inspiration from the data augmentation scheme employed by [54, 44], we augment the radar data in a similar fashion. However, we train the radar model on both the unsegmented and augmented data simultaneously, and further with an MSE loss to ensure the data augmentation produces the exact same waveform as the unsegmented data.

Fusion Network: The fusion is conducted in 2 steps. First, we mimic the rPPG-fusion network to fuse the 2 thermal cameras. However, the thermal-fusion network is trained from scratch without pretraining. Following the thermal-fusion, we adopt a late fusion strategy to fuse the radar waveforms with the thermal-fusion waveforms. For this waveform fusion network, a 1-D CNN is trained on both remote respiratory waveforms concatenated along the channel dimension. This approach accounts for the fundamental similarities between camera modalities and the differences between cameras and radar. Hence, mid-fusion networks for cameras and late-fusion for camera+radar modalities are followed.

4.3 Data Loading and Processing

Preprocessing Details:

- Video spatial resolution: The 256×256 frames are downsized to 128×128 for the model
- Downsampling: cardiac-related modalities are downsampled to 30 Hz, while respiratory-related to 15 Hz.
- NIR and thermal videos: channels duplicated $3\times$ to match RGB input shape

Window Lengths by Task:

- Waveform regression training: 10 sec (rPPG) / 20 sec (respiratory)
- Vital sign calculation: 30 sec windows (HR) / 40 sec windows (RR)
- Cognitive load feature extraction: Full 2 min recordings

4.4 Resources and Hyperparameters

Here we provide a brief summary of the hyperparameters used. Due to the number of baselines and methods we ran, the hyperparameter tuning was very brief. An exhaustive

grid search would yield models with errors lower than those tabulated in the main paper. Here, we consolidate the important hyperparameters used for each model in the table below. Detailed values for each model can be found in the config files in our repository: https://github.com/AnirudhBHarish/CogPhys/tree/main/configs/train_configs.

Table 3: **Comparison of Methods and Hyperparameters.** The rPPG models and resp models were trained for a different number of epochs. Hence, we represent it as rppg-epochs/resp-epochs. The input length here is represented in samples. These correspond to 10-second and 20-second windows for rPPG and remote resp, respectively.

Method	**Hyperparameters**					
	LR	**Epochs**	**Opt**	**Batch**	**In Len**	**Loss**
PhysNet	1e-5	50/100	Adam	2	300	NegPearson + SNR
RhythmFormer	9e-3	30/30	AdamW	1	300	RhythmFormer Loss
PhysFormer	1e-4	50/100	Adam	2	300	PhysFormer Loss
FactorizePhys	1e-5	50/100	Adam	2	300	Smooth NegPearson2 + SNR
PhysMamba	1e-5	50/100	Adam	2	300	Smooth NegPearson2 + SNR
Contrast-Phys+	1e-5	50/100	AdamW	2	300	ConstrastLoss + Smooth NegPearson2 + SNR
Fusion Contrast-Phys+	1e-5	10/100	AdamW	2	300	ConstrastLoss + Smooth NegPearson2 + SNR
RF-Net	1e-3	10	AdamW	2	300	ConstrastLoss + Smooth NegPearson + SNR + MSE
Waveform Fusion	1e-3	50	AdamW	32	300	ConstrastLoss + Smooth NegPearson2 + SNR

In reference to Table 3, we note the following:

- Since the rPPG fusion model is pre-trained, we only train it for 10 epochs.
- RhythmFormer has very different hyperparameters compared to the rest of the models due to its significantly long computation time.
- Refer to the original ContrastPhys+ [41], PhysFormer [52], and RhythmFormer [55] papers for additional information regarding ContrastLoss, PhysFormer Loss, and RhythmFormer Loss, respectively.

The following adjustments were made to the loss functions:

- Contrastive Loss was modified. The error/distance metric was swapped from mean squared error (*MSE*) to a combined loss: *MSE + Kullback–Leibler (KL) divergence*
- The *SNRLoss* [44], i.e., a frequency loss was included in training of the convolutional models.
- We create a variant of the NegPerson loss called the Smooth NegPearson2 loss. The smooth here refers to an additional smoothing operator (box-flter) applied to the signals before computing the loss. The 2 refers to the expotent term to which we loss is taken. That is

$$\textit{Smooth NegPearson} = \textit{NegPearson}(w \circledast \text{pred}, w \circledast \text{gt}), \tag{1}$$

$$\textit{Smooth NegPearson2} = (\textit{NegPearson}(w \circledast \text{pred}, w \circledast \text{gt}))^2. \tag{2}$$

Hyperparmeter tuning was to primarily decide the addition of the extra loss functions and preprocessing techniques. That is, the *rppg-toolbox* paper and codebase include frame-differencing as a preprocessing step. Our early stage hyperparameter tuning did not yield favorable results with this preprocessing step. Hence, we dropped the same in our subsequent iterations. The model weights have been released for reproducibility

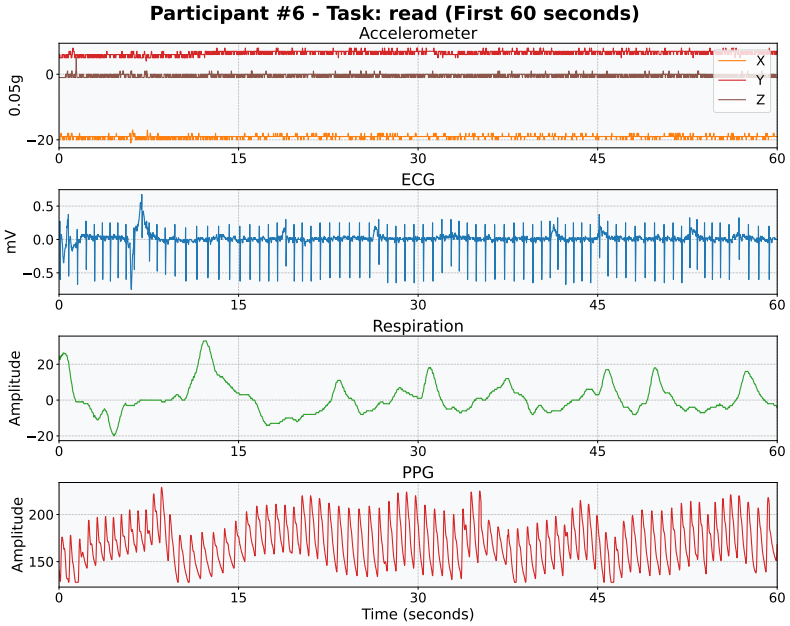


Figure 3: Example of vital signs waveforms recorded by contact sensors

All our remote-vitals sign estimation models were trained on servers with *NVIDIA GeForce RTX 2080* GPUs. Each GPU has 12 GB of compute memory. All the unimodal methods and waveform fusion methods were trained with a single GPU, while the camera-fusion method was trained with 2 of the GPUs listed above. The dataset was also hosted on the server and occupied ~ 611 GBs of memory. Upon preprocessing the data, the chunked data occupies ~ 270 GBs.

5 Estimating Cognitive Load

5.1 ML Framework for Cognitive Load Classification

We developed a benchmarking framework to evaluate cognitive load classification using physiological signals from multiple modalities. Our approach encompasses both traditional ML classifiers operating on engineered features and DL architectures that learn representations directly from raw signals. The feature extraction pipeline incorporates both time-domain and frequency-domain analyses following established methods in physiological signal processing. An example of raw signal waveforms from the contact sensors is showcased in Figure 3.

5.1.1 Feature Engineering Pipeline

Our feature extraction pipeline transforms raw physiological signals into features that capture cardiovascular, respiratory, and ocular dynamics associated with cognitive load. The pipeline processes signals from multiple modalities with modality-specific preprocessing and feature extraction methods.

PPG Signal Processing and Feature Extraction Raw PPG signals undergo preprocessing with a third-order Butterworth bandpass filter $(0.8 - 3.0\ Hz)$ to isolate the cardiac frequency range, $48 - 180$ beats-per-minute (BPM), while attenuating baseline wander and high-frequency noise. We perform robust peak detection on the filtered waveform using physiological constraints: minimum inter-peak distance of 0.5 $secs$ and adaptive height thresholds based on signal characteristics. From the detected peaks, we calculate inter-beat intervals (IBIs) and filter out physiologically implausible values (< 0.3 $secs$ or > 1.5 $secs$, corresponding to HR values outside $40 - 200BPM$). Figure 4 shows two examples of 30 seconds of filtered PPG signals and the extracted peaks. Figure 5 shows the differences in average HR across tasks, potentially caused by changes in cognitive load level.

From the processed PPG signals, we extract a set of 25+ features encompassing:

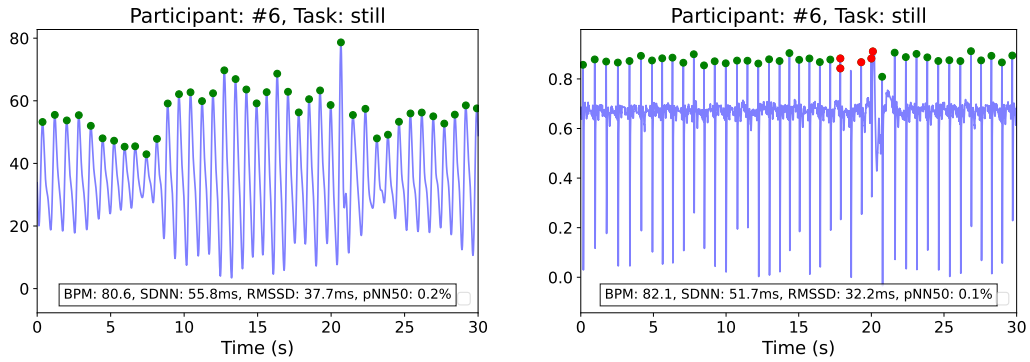


Figure 4: Comparison of filtered PPG and ECG waveforms. (Left) Filtered PPG waveform with identified peaks (marked as green dots) and calculated features. (Right) Filtered ECG waveform with identified peaks and calculated features.

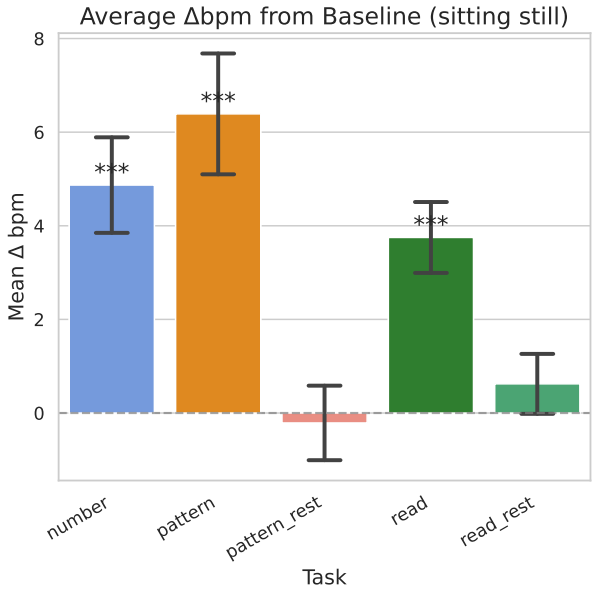


Figure 5: Task-induced changes in HR (BPM) compared to the baseline "still" condition. The bars represent mean differences, with standard error bars indicating measurement variability, while asterisks "***" denote statistical significance p<0.001 of the difference from baseline as determined by one-sample t-tests.

- *Time-domain HRV features*: mean HR (BPM), RMSSD (root mean square of successive differences), SDNN (standard deviation of NN intervals), pNN50 (proportion of successive intervals differing by $> 50ms$), and pNN20
- *Poincaré plot features*: SD1 (short-term variability), SD2 (long-term variability), and SD1/SD2 ratio reflecting the balance between short- and long-term cardiac variability
- *Frequency-domain features*: VLF power $(0.0033-0.04\ Hz)$, LF power$(0.04-0.15\ Hz)$, HF power $(0.15-0.4\ Hz)$, and LF/HF ratio—a key indicator of sympathetic/parasympathetic balance
- *Pulse morphology features*: pulse amplitude statistics (mean, standard deviation) and pulse width metrics derived from half-amplitude measurements
- *Statistical features*: signal mean, standard deviation, range, and spectral characteristics in the cardiac frequency band

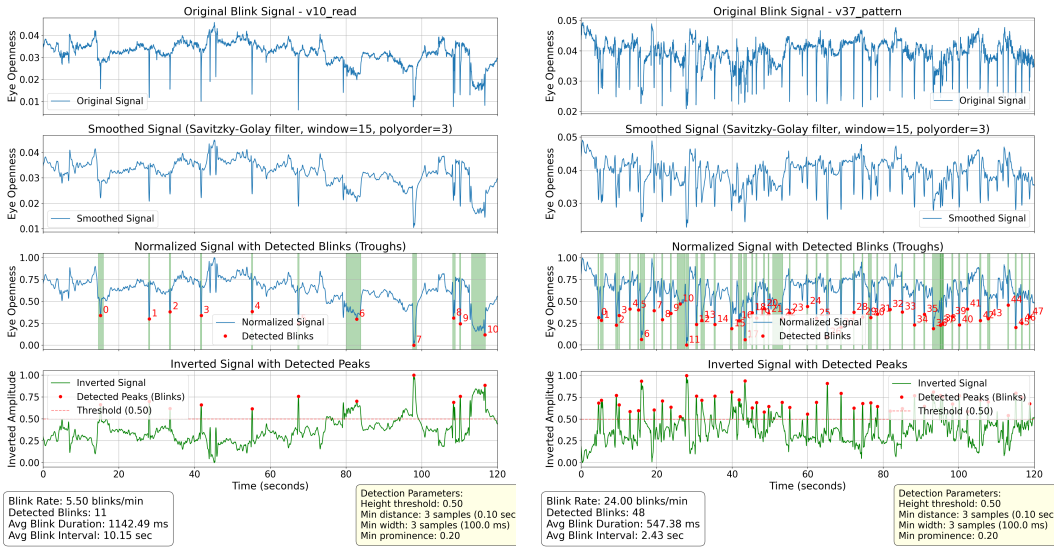


Figure 6: Eye blinking detection from normalized eye-openness 2 min waveforms derived from camera frames. Left and right subfigures are from two participants with different blinking patterns. Rows (1) Raw eye-openness signal. (2) Signals smoothed by Savitzky-Golay filter. (3) Blinks detected are marked as red dots. (4) The inverted signal used for peak detection.

Respiratory Signal Processing For respiratory signals, we apply a second-order Butterworth bandpass filter $(0.05 - 1.0\ Hz)$ corresponding to $3 - 60\ RPM$ (resps per minute). We extract RR through spectral analysis, identifying the dominant frequency in the respiratory band. Additional features include respiratory power, spectral entropy as a measure of breathing pattern complexity, and basic statistical descriptors of the respiratory waveform.

Blink Signal Processing Eye openness signals are processed to detect blink events by identifying troughs (minima) in the normalized eye-openness signal, as blinks correspond to reduced eye aperture. We apply Savitzky-Golay smoothing (window=15, polynomial order=3) and invert the normalized signal to convert troughs to peaks for detection. Extracted features include blink rate, blink duration statistics (mean, standard deviation), inter-blink interval variability, blink amplitude characteristics, and blink depth measurements. Examples of detected eye blinking are shown in Figure 6.

5.1.2 Mathematical Formulations of PPG Features for ML models

The key physiological features are computed using the following mathematical formulations:

$$HR = \frac{60}{\text{mean(IBI)}}, \quad RR = \frac{60}{\text{mean(breath intervals)}} \tag{3}$$

$$\text{RMSSD} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(IBI_{i+1} - IBI_i)^2}, \quad \text{SDNN} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(IBI_i - \bar{IBI})^2} \tag{4}$$

$$S(f) = \frac{1}{N}\left|\sum_{i=1}^{N} IBI_i e^{-j2\pi f t_i}\right|^2, \quad \text{LF/HF} = \frac{\int_{0.04}^{0.15} S(f)df}{\int_{0.15}^{0.4} S(f)df} \tag{5}$$

5.1.3 ML Models

We evaluated eight traditional ML classifiers, each configured with optimized hyperparameters and class balancing strategies:

- *Random Forest (RF)*: 200 estimators, balanced class weights, unlimited depth

- *Gradient Boosting (GB)*: 200 estimators, learning rate 0.05, max depth 4, subsample 0.8
- *Support Vector Machine (SVM)*: RBF kernel, balanced class weights, probability estimates enabled
- *Logistic Regression (LR)*: L2 regularization, liblinear solver, balanced class weights
- *Linear Discriminant Analysis (LDA)*: SVD solver with covariance storage
- *K-Nearest Neighbors (KNN)*: 5 neighbors, distance weighting, Euclidean metric
- *Decision Tree (DT)*: Gini criterion, balanced class weights, unlimited depth
- *Multi-Layer Perceptron (MLP)*: Single hidden layer (100 units), ReLU activation, Adam optimizer

All models operate on standardized features using StandardScaler preprocessing within scikit-learn pipelines to ensure consistent feature scaling across modalities.

5.1.4 DL Architectures

We implemented three DL architectures that process raw physiological signals directly, eliminating the need for manual feature engineering:

1D CNN A multi-layer 1D CNN with increasing filter sizes across convolutional blocks, batch normalization, ReLU activations, and dropout for regularization. The architecture processes multi-channel time series data with adaptive input channel handling.

Long Short-Term Memory (LSTM) A recurrent architecture with 2 LSTM layers (128 hidden units each) designed to capture temporal dependencies in physiological signals. The model processes sequential data and outputs class probabilities through a fully connected layer.

ResNet1D A 1D adaptation of the ResNet architecture incorporating residual connections to enable training of deeper networks. Multiple residual blocks with skip connections facilitate gradient flow and feature learning at different temporal scales.

All DL models use a learning rate of 0.0005, batch size of 16, and are trained for a maximum of 20 epochs with early stopping (patience=10) based on validation loss. We employ robust data preprocessing, including NaN interpolation and per-channel robust normalization using median and interquartile range scaling.

5.1.5 Experimental Configurations

We systematically evaluated seven distinct physiological signal combinations to investigate the utility of different sensing modalities:

1. *Contact PPG Only*: High-fidelity pulse oximeter data as baseline
2. *rPPG Only*: Camera-derived rPPG signals
3. *Blink Markers Only*: Eye openness and blink dynamics
4. *Contact PPG + Contact Respiratory*: Multimodal contact sensing
5. *rPPG + Contact Respiratory*: Hybrid contact-remote approach
6. *rPPG + Remote Respiratory*: Fully remote cardiorespiratory sensing
7. *rPPG + Remote Respiratory + Blink Markers*: Complete multimodal remote sensing
8. *Contact PPG + Contact Respiratory + Blink Markers*: Vitals from contact sensors and blink, for comparison with 7. and to assess the remote vitals sensing quality

This experimental design enables systematic comparison of contact versus remote sensing effectiveness, quantification of multimodal integration benefits, and assessment of the relative contributions of different physiological indicators to cognitive load detection.

5.1.6 Training and Evaluation Methodology

The dataset was partitioned using participant-based splits to ensure generalization to unseen individuals (consistent with the partitioning used in the remote sensing stage): 25 participants for training, 2 for validation, and 10 for testing. This split strategy prevents data leakage and provides realistic performance estimates for deployment scenarios where models must work on new users without personalized calibration.

For ML models, hyperparameter optimization was performed on the validation set, while DL models utilized validation data for early stopping to prevent overfitting. All models were evaluated on the held-out test set using multiple metrics: accuracy, weighted F1-score, precision, recall, and specificity. We generated confusion matrices to analyze classification patterns and computed class-specific performance metrics to account for potential label imbalance.

Signal preprocessing included robust handling of missing data through linear interpolation, application of modality-specific filtering (bandpass filters for cardiac and respiratory signals, smoothing for blink signals), and normalization strategies adapted to each signal type. For multimodal experiments, signals were temporally aligned and combined into multi-channel arrays with consistent sampling rates and durations (120 seconds per task).

5.1.7 Ablation Study: Mental Demand Only Labeling

To address concerns about potential confounding between stress and cognitive load in our composite NASA-TLX score, we conducted an ablation study using only the "Mental Demand" subscale for binary labeling. The median-split approach was applied to the Mental Demand scores alone (rather than the composite of Mental Demand, Temporal Demand, Effort, and Frustration). The results demonstrate remarkable consistency: the original 4-dimension approach yielded 76 Low/76 High labels (perfectly balanced), while the Mental Demand only approach yielded 75 Low/77 High labels (nearly identical distribution, with only 1 sample differing from 152 total). Performance evaluation using rPPG + Remote Resp + Blink Markers with Gradient Boosting showed 86.49% accuracy (F1: 0.878) for the original approach versus 81.08% accuracy (F1: 0.837) for Mental Demand only. The small performance difference ($\sim 5\%$) is within the range of training variability and likely reflects the single differing label rather than measuring fundamentally different physiological constructs. This validates the robustness of our original multi-dimensional labeling approach while confirming that mental demand remains the primary driver of our cognitive load classifications.

6 Remote Vital Sensing Benchmarking Extended Results

We analyze the best unimodal and multimodal networks for HR and RR estimation through Figure 7. Here, Contrast-Phys+ was the best-performing model for the NIR, RGB, and thermal modalities. For rPPG estimation, we see that the fusion algorithm outperforms both the RGB and NIR methods, excelling at preserving frequency content and the morphological/shape-based properties. While the RGB does retain frequency information, it loses morphological information of the PPG signals as seen in the second example. For RR estimation, we can see that the radar and waveform fusion methods reliably retained the waveform's structure across samples, with thermal below performing well for the sample below. Most notably, we can see that the respiratory GT signals itself is prone to aperiodic trends, causing benchmarking of RR performance to be challenging.

6.1 Train Configuration

Both rPPG and the remote respiratory models were trained on 300-sample length signals. This equates to $10\ secs$ of rPPG data and $20\ secs$ of respiratory data at 30 and $15\ Hz$, respectively. Furthermore, data from the NIR and thermal cameras were duplicated along the channel dimension three times to match the shape of the RGB videos. When using the Negative Person loss function, the waveforms were smoothed with a box filter of sizes 7 and 15 for rPPG and respiratory signals, respectively.

Our dataset and code base include specific hardware details, such as sensor failure, flatlining, and sample dropping due to corrupted sensor recordings.

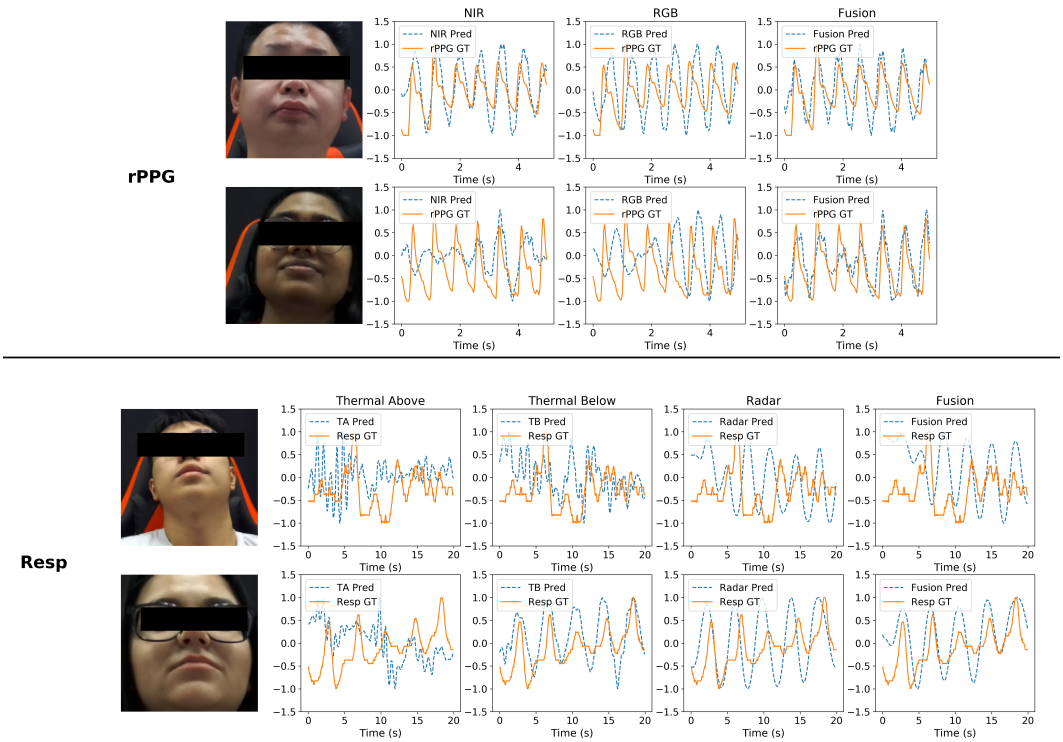


Figure 7: **Comparison of Sample Output Waveforms for Different Modalities.** For all the cameras, we plot the waveforms from the Contrast-Phys+ models, while the RF-Net and Waveform Fusion were chosen for the radar and respiratory fusion. The rPPG-fusion algorithm is superior to both RGB and NIR methods for rPPG estimation. For RR estimation, the waveform fusion and radar networks reliably preserve the waveform structure across samples.

6.2 Test Configuration

We take non-overlapping 30 $secs$ (900 samples at 30 Hz) duration windows to calculate the HR and HRV values for the estimated rPPG and ground truth PPG signals. Similarly, we consider non-overlapping 40 $secs$ (600 samples at 15 Hz) when calculating RR from remote and content respiratory signals. Following [44], we report the mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and the Pearson correlation coefficient (r) for HR and RR values. For HRV, we report the MAE values for IBI.

To calculate the HR values, the extracted waveforms are normalized, detrended, and filtered with box-filters and 6^{th} order Butterworth filters, with cutoff frequencies at 40 and 180 BPM. A Welch-based periodogram method is then used to calculate the HR values. Similarly, to calculate RR, we box-filter, detrend, and filter the resultant signal with a 2^{nd} order filter, with cutoff frequencies at 8 and 30 BPM. However, unlike HR estimation, we do not employ the Welch's method and resort to a simpler periodogram implementation. In addition to HR, we also calculate IBI. For this, we apply a second box filter to the rPPG waveform, followed by the peak detection algorithm from the *heartpy* library [42]. Further details can be found in the codebase.

6.3 Waveform Metrics and Clinical Significance (Main Fold only)

Here, we introduce two waveform metrics, namely SNR and Mean Absolute Cross-Correlation (MACC). We also introduce a clinical significance metric introduced by the Association for the Advancement of Medical Instrumentation (AAMI) for cardiac monitoring.

Note: The values in this subsection are for the main fold from the main paper. The cross-validation values are presented in the next subsection.

Table 4: **Performance of standard rPPG algorithms on the main fold *CogPhys* dataset (continuation of the metrics in the main paper).** e report signal-to-noise ratio (SNR), mean absolute cross-correlation (MACC), and clinical performance according to AAMI standards. The standard error spread for each metric has also been tabulated. 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.

		Waveform Metrics		Clinical Metric
	Method	SNR (dB) ↑	MACC ↑	AAMI (%) ↑
NIR	**PhysNet** [50]	-5.93 ± 0.18	0.17 ± 0.00	25.00
	RythmFormer [55]	-1.66 ± 0.49	0.51 ± 0.01	48.73
	PhysFormer [52]	-8.01 ± 0.81	0.33 ± 0.01	31.36
	FactorizePhys [18]	-3.37 ± 0.33	0.48 ± 0.01	52.12
	PhysMamba [27]	-1.64 ± 0.29	0.53 ± 0.01	70.34
	Contrast-Phys+ [41]	-1.04 ± 0.33	$\underline{0.56} \pm 0.01$	66.10
	Pretrained Contrast-Phys+	-0.59 ± 0.33	$\mathbf{0.58} \pm 0.01$	72.03
RGB	**PhysNet** [50]	-5.87 ± 0.16	0.15 ± 0.00	27.12
	RythmFormer [55]	-1.27 ± 0.52	0.42 ± 0.01	50.00
	PhysFormer [52]	0.14 ± 0.68	0.43 ± 0.01	54.66
	FactorizePhys [18]	-0.61 ± 0.31	0.45 ± 0.01	76.27
	PhysMamba [27]	0.15 ± 0.27	0.47 ± 0.01	$\underline{80.51}$
	Contrast-Phys+ [41]	$\underline{0.73} \pm 0.31$	0.48 ± 0.01	79.66
	Fusion	$\mathbf{1.20} \pm 0.29$	0.49 ± 0.01	**85.17**

We evaluate waveform quality and clinical viability using three complementary metrics. SNR quantifies the power ratio between the extracted physiological signal and background noise, expressed in decibels (dB), where higher values indicate cleaner signal extraction with less contamination from noise sources. MACC measures morphological similarity between predicted and ground-truth waveforms by computing the average peak cross-correlation across temporal windows, ranging from 0 to 1, where values closer to 1 indicate better preservation of the cardiac or respiratory waveform shape and timing [8]. AAMI Clinical Performance assesses clinical acceptability according to the Association for the Advancement of Medical Instrumentation EC13:2002 standard for cardiac monitors, which requires that HR estimates fall within the larger of $\pm 5\ BPM$ or $\pm 10\%$ of the ground truth. The metric reports the percentage of estimates meeting this criterion [13].

From Table 6.3, we observe that the Fusion delivers the strongest overall results with 85.17% AAMI compliance and the highest SNR. When analyzing MACC on the other hand, the NIR methods consistently dominate MACC scores—with the top three all from NIR—while RGB methods show a clear advantage in clinical metrics, with three algorithms exceeding 75% AAMI versus NIR's maximum of 72%. The SNR pattern reinforces this split, as RGB produces higher values with several methods reaching positive SNR, while all NIR results remain negative. This suggests NIR excels at preserving temporal alignment while RGB provides cleaner signals that translate more effectively to clinical performance. The fusion leverages both modalities' strengths: its MACC (0.49) sits between RGB's ceiling (0.48) and NIR's peak (0.58), while its SNR $(1.20\ dB)$ surpasses both modalities, translating to a 5-13 percentage point AAMI gain over the best single-modality results (85.17% vs 80.51% RGB, 72.03% NIR).

From Table 5, RF-Net achieves the highest SNR at $6.34\ dB$-nearly double the best camera result—yet its MACC of 0.54 is exceeded by camera methods, with Contrast-Phys+ for the thermal placed below reaching 0.59. The thermal camera positioned below consistently outperforms the one positioned above across all algorithms, with MACC improvements of 0.06-0.15 points, suggesting this angle captures respiratory movement more effectively. The fusion strategies diverge: Camera Fusion achieves 0.57 MACC with moderate SNR $(3.40\ dB)$, while Waveform Fusion prioritizes SNR $(3.67\ dB)$ over correlation (0.52)

6.4 Cross-Validation and Clinical Significance Analysis (All Folds)

To provide robust performance estimates, we conducted 4-fold cross-validation for the top-performing HR and RR estimation methods. All 37 participants appeared in the test set exactly once across the 4 folds, with test set sizes of {10, 9, 9, 9} participants. Note that this is different than the train/valid/test split used in the main paper, which was described in section 5.1.6. Please note, the RythmFormer

Table 5: **Performance of resp. rate (RR) estimation algorithms on the main fold *CogPhys* dataset (continuation of the metrics in the main paper).** We report the SNRMACC. The standard error spread for each metric has also been tabulated. 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.

		Waveform Metrics	
	Method	SNR (dB) ↑	MACC ↑
Above	**PhysNet** [50]	0.29 ± 0.40	0.38 ± 0.01
	RythmFormer [55]	2.27 ± 0.27	0.42 ± 0.00
	PhysFormer [52]	2.63 ± 0.92	0.44 ± 0.01
	FactorizePhys []	3.09 ± 0.37	0.45 ± 0.01
	PhysMamba [27]	1.12 ± 0.38	0.41 ± 0.01
	Contrast-Phys+ [41]	1.33 ± 0.48	0.44 ± 0.01
Below	**PhysNet** [50]	2.90 ± 0.55	0.53 ± 0.01
	RythmFormer [55]	3.56 ± 0.57	0.54 ± 0.01
	PhysFormer [52]	1.33 ± 0.94	0.47 ± 0.01
	FactorizePhys [18]	$\underline{3.87} \pm 0.45$	0.54 ± 0.01
	PhysMamba [27]	3.00 ± 0.55	$\underline{0.57} \pm 0.01$
	Contrast-Phys+ [41]	3.61 ± 0.60	$\mathbf{0.59} \pm 0.01$
RF	**RF-Net**	$\mathbf{6.34} \pm 0.59$	0.54 ± 0.01
	Cameras Fusion	3.40 ± 0.58	$\underline{0.57} \pm 0.01$
	Waveform Fusion	3.67 ± 0.66	0.52 ± 0.01

Table 6: **Performance of HR estimation algorithms on the *CogPhys* dataset (all folds** We report the mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), Pearson correlation (r), and inter-beat interval (IBI) error. 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.

		HR Metrics (BPM)				HRV Metric
	Method	MAE ↓	RMSE ↓	MAPE ↓	r ↑	IBI (ms) ↓
NIR	**PhysNet** [50]	13.58 ± 0.34	16.93 ± 3.55	17.46 ± 0.49	-0.05 ± 0.03	320.95 ± 7.11
	PhysFormer [52]	12.28 ± 0.33	15.75 ± 3.62	16.22 ± 0.51	-0.01 ± 0.03	129.36 ± 4.83
	FactorizePhys [18]	9.37 ± 0.44	15.95 ± 4.19	10.95 ± 0.48	0.40 ± 0.03	82.96 ± 3.00
	PhysMamba [27]	5.76 ± 0.31	10.71 ± 3.20	7.89 ± 0.50	0.58 ± 0.03	72.46 ± 3.49
	Contrast-Phys+ [41]	7.80 ± 0.34	12.69 ± 3.40	10.38 ± 0.53	0.37 ± 0.03	74.41 ± 3.37
	Pretrained Contrast-Phys+	5.83 ± 0.31	10.88 ± 3.28	7.90 ± 0.51	0.55 ± 0.03	68.23 ± 3.36
RGB	**Green** [43]	22.91 ± 0.53	27.68 ± 5.17	27.32 ± 0.56	0.01 ± 0.03	119.74 ± 3.15
	ICA [36]	14.08 ± 0.50	20.48 ± 4.67	16.38 ± 0.54	0.20 ± 0.03	74.09 ± 2.60
	CHROM [10]	5.63 ± 0.29	10.23 ± 3.12	6.67 ± 0.32	0.68 ± 0.02	58.26 ± 2.82
	POS [45]	5.33 ± 0.35	11.73 ± 4.79	6.52 ± 0.43	0.63 ± 0.03	50.66 ± 2.63
	PhysNet [50]	11.27 ± 0.34	15.17 ± 3.44	14.48 ± 0.48	0.11 ± 0.03	274.93 ± 7.17
	PhysFormer [52]	10.12 ± 0.37	14.96 ± 3.89	13.17 ± 0.53	0.13 ± 0.03	97.09 ± 3.96
	FactorizePhys [18]	4.62 ± 0.32	10.54 ± 3.73	$\mathbf{5.30} \pm 0.34$	0.69 ± 0.02	$\mathbf{40.04} \pm 1.66$
	PhysMamba [27]	$\mathbf{4.36} \pm 0.27$	$\mathbf{9.14} \pm 3.21$	$\underline{5.36} \pm 0.34$	$\mathbf{0.72} \pm 0.02$	48.77 ± 2.78
	Contrast-Phys+ [41]	4.60 ± 0.27	9.32 ± 3.06	6.16 ± 0.44	$\underline{0.69} \pm 0.02$	50.05 ± 3.09
	Fusion	$\underline{4.45} \pm 0.28$	$\underline{9.32} \pm 3.16$	6.07 ± 0.46	$\underline{0.69} \pm 0.02$	$\underline{47.87} \pm 3.18$

model was found to be extremely computationally intensive. Hence, cross-fold training of the same was unfeasible.

6.4.1 HR Cross Validation and Significance Testing

Based on the tabulated values from Tables 6 & 7, PhysMamba RGB achieves the best HR estimation performance, with the lowest MAE $= 4.36\ BPM$, albeit marginally. The RGB-NIR fusion approach emerges as the second best, trailing PhysMamba by only 2% in MAE, while delivering a substantially superior SNR. We further test this via the ANOVA and Tukey HSD pairwise testing. Specifically, we run the tests on the best performing RGB, NIR, and Fusion methods, i.e., PhysMamba (RGB), pretrained and finetuned ContrastPhys+ (NIR), and the Siamese Fusion Network (Fusion).

ANOVA confirms significant performance differences across methods ($F = 8.06, p < 0.001$), with post-hoc analysis revealing that both Fusion and RGB significantly outperform NIR ($p = 0.002$ vs

Table 7: **Waveform and clinical metrics for HR estimation on the *CogPhys* dataset (all folds).** We report SNR, MACC, and clinical performance according to AAMI standards. 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.

		Waveform Metrics		Clinical Metric
	Method	SNR (dB) ↑	MACC ↑	AAMI (%) ↑
NIR	**PhysNet** [50]	-6.01 ± 0.09	0.17 ± 0.00	23.85
	PhysFormer [52]	-8.54 ± 0.39	0.30 ± 0.01	26.83
	FactorizePhys [18]	-2.83 ± 0.17	0.50 ± 0.01	60.67
	PhysMamba [27]	-1.06 ± 0.15	$\underline{0.55} \pm 0.01$	70.87
	Contrast-Phys+ [41]	-1.87 ± 0.17	0.51 ± 0.01	57.91
	Pretrained Contrast-Phys+	-0.41 ± 0.17	$\mathbf{0.58} \pm 0.01$	69.38
RGB	**Green** [43]	-8.38 ± 0.13	0.24 ± 0.00	19.41
	ICA [36]	-4.17 ± 0.18	0.37 ± 0.00	44.86
	CHROM [10]	-2.36 ± 0.13	0.42 ± 0.00	71.58
	POS [45]	-1.47 ± 0.14	0.44 ± 0.00	75.11
	PhysNet [50]	-4.83 ± 0.12	0.20 ± 0.00	37.21
	PhysFormer [52]	-3.13 ± 0.38	0.38 ± 0.01	44.63
	FactorizePhys [18]	-0.00 ± 0.15	0.47 ± 0.00	**81.62**
	PhysMamba [27]	0.44 ± 0.15	0.48 ± 0.00	$\underline{80.37}$
	Contrast-Phys+ [41]	$\underline{1.13} \pm 0.16$	0.50 ± 0.00	79.34
	Fusion	$\mathbf{1.35} \pm 0.15$	0.50 ± 0.00	$\underline{80.73}$

Table 8: **Performance of RR estimation algorithms on the *CogPhys* dataset (all folds).** We report the MAE, RMSE, MAPE, Pearson correlation (r), SNR, and MACC. The best and second-best performing numbers are shown in **bold** and underline, respectively.

		RR Metrics (RPM)				Waveform Metrics	
	Method	MAE ↓	RMSE ↓	MAPE ↓	r ↑	SNR (dB) ↑	MACC ↑
Above	**PhysNet** [50]	3.36 ± 0.11	4.30 ± 1.02	19.41 ± 0.67	0.02 ± 0.04	-0.89 ± 0.22	0.37 ± 0.00
	PhysFormer [52]	4.07 ± 0.12	5.11 ± 1.18	21.96 ± 0.61	-0.01 ± 0.04	-0.67 ± 0.45	0.42 ± 0.01
	FactorizePhys [18]	3.06 ± 0.11	4.19 ± 1.06	18.44 ± 0.79	0.08 ± 0.04	2.32 ± 0.19	0.44 ± 0.00
	PhysMamba [27]	5.07 ± 0.15	6.39 ± 1.43	27.67 ± 0.79	-0.01 ± 0.04	0.61 ± 0.16	0.40 ± 0.00
	Contrast-Phys+ [41]	2.45 ± 0.10	$\underline{3.47} \pm 0.89$	14.24 ± 0.61	0.25 ± 0.04	1.18 ± 0.25	0.46 ± 0.01
Below	**PhysNet** [50]	2.57 ± 0.11	3.74 ± 1.01	14.55 ± 0.62	0.24 ± 0.04	1.11 ± 0.28	0.50 ± 0.01
	PhysFormer [52]	3.64 ± 0.13	4.84 ± 1.14	19.49 ± 0.63	0.14 ± 0.04	-1.07 ± 0.45	0.46 ± 0.01
	FactorizePhys [18]	2.40 ± 0.11	3.70 ± 1.14	14.28 ± 0.72	0.33 ± 0.04	$\underline{3.48} \pm 0.23$	0.54 ± 0.01
	PhysMamba [27]	$\mathbf{2.25} \pm 0.10$	$\underline{3.45} \pm 0.95$	$\underline{12.83} \pm 0.61$	$\underline{0.34} \pm 0.04$	2.58 ± 0.29	$\underline{0.57} \pm 0.01$
	Contrast-Phys+ [41]	$\mathbf{2.25} \pm 0.11$	3.54 ± 1.09	$\mathbf{12.66} \pm 0.62$	$\mathbf{0.38} \pm 0.04$	2.22 ± 0.31	$\mathbf{0.58} \pm 0.01$
RF	**RF-Net**	2.45 ± 0.10	$\underline{3.45} \pm 0.91$	13.73 ± 0.55	0.24 ± 0.04	$\mathbf{4.09} \pm 0.32$	0.55 ± 0.01
	Cameras Fusion	$\underline{2.29} \pm 0.10$	$\mathbf{3.43} \pm 0.93$	$\mathbf{12.80} \pm 0.57$	0.32 ± 0.04	1.80 ± 0.31	0.56 ± 0.01
	Waveform Fusion	2.52 ± 0.10	3.62 ± 1.02	13.85 ± 0.53	0.23 ± 0.04	1.73 ± 0.35	0.50 ± 0.01

Fusion and $p = 0.001$ vs RGB), while showing no significant difference from each other ($p = 0.98$). RGB and Fusion also achieve equally high clinical compliance of 80.73%.

6.4.2 RR Cross Validation and Significance Testing

Table 8 reveals that RR estimation exhibits more uniform performance across methods compared to HR. Thermal Below Contrast-Phys+ and Thermal Below PhysMamba share the top position with identical MAE values of $2.25 RPM$, closely followed by the Cameras Fusion approach at $2.29\ RPM$. Notably, Radar achieves superior waveform quality with the highest SNR of $4.09\ dB$. To assess statistical significance, we conduct ANOVA and Tukey HSD pairwise testing across Cameras Fusion, Radar, Waveform Fusion, and Thermal Below CP+.

The analysis demonstrates no statistically significant differences among methods (F=1.50, p=0.214), with all pairwise comparisons yielding $p > 0.25$. The post-hoc analysis further confirmed no significant pairwise differences between Cameras Fusion ($MAE = 2.29\ RPM$), Radar ($MAE = 2.45\ RPM$), Waveform Fusion ($MAE = 2.52\ RPM$), and Thermal Below CP+ ($MAE = 2.25\ RPM$, all $p > 0.25$). All methods achieve statistically equivalent accuracy for RR estimation.

Table 9: 4-fold cross-validation performance comparison of ML models across seven physiological signal combinations for mental workload classification. Showing Accuracy(F1 Score) in each cell. **Bold** values indicate the highest accuracy for each model across all experimental configurations; underlined values indicate the second-highest accuracy. The overall best accuracy 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
RF	0.65(0.65)	0.54(0.56)	0.53(0.56)	0.68(0.68)	**0.81(0.81)**	0.54(0.56)	0.60(0.62)	0.80(0.82)
GB	0.60(0.59)	0.55(0.57)	0.57(0.59)	0.64(0.63)	**0.82(0.82)**	0.55(0.57)	0.61(0.63)	0.78(0.80)
SVM	0.60(0.60)	0.62(0.68)	0.55(0.61)	0.62(0.64)	0.83(0.84)	0.62(0.68)	0.59(0.63)	**0.83(0.85)**
LR	0.62(0.63)	0.59(0.59)	0.52(0.56)	0.64(0.64)	0.79(0.80)	0.59(0.59)	0.58(0.56)	**0.81(0.83)**
LDA	0.62(0.63)	0.59(0.60)	0.51(0.53)	0.58(0.60)	**0.77(0.78)**	0.59(0.60)	0.58(0.60)	0.73(0.75)
KNN	0.53(0.56)	0.54(0.58)	0.52(0.62)	0.56(0.62)	**0.72(0.76)**	0.54(0.58)	0.58(0.62)	0.66(0.73)
DT	0.51(0.44)	0.45(0.37)	0.52(0.58)	0.62(0.59)	0.64(0.64)	0.45(0.37)	0.50(0.54)	**0.66(0.69)**
MLP	0.62(0.62)	0.62(0.62)	0.57(0.61)	0.62(0.61)	**0.84(0.84)**	0.62(0.62)	0.63(0.65)	0.80(0.82)

6.4.3 Cognitive Load classification Cross-Validation

The 4-fold cross-validation results for cognitive load classification, detailed in Table 9, reveal several key trends regarding model and modality performance. The primary insight is the significant performance gain achieved through multimodal signal integration. While unimodal approaches using either remote or contact PPG yield only moderate accuracies, their fusion with respiratory and, critically, ocular data leads to substantial improvements across all models. Notably, the combination of fully remote signals (rPPG, Remote Respiratory, and Blink Markers) consistently outperforms any single modality, including the high-fidelity contact PPG sensor. For example, the SVM model reaches 83%accuracy with the full remote fusion, far surpassing the 65% accuracy from the top model using only contact PPG. This finding strongly validates the central hypothesis that fusing multiple, lower-quality remote signals can yield a more informative representation for cognitive state estimation than a single, high-quality contact signal. Moreover, the performance of this fully remote configuration (83% accuracy, 0.85 F1) closely approaches that of its contact-based counterpart (84% accuracy, 0.84 F1), demonstrating the increasing viability of non-intrusive sensing for reliable cognitive load assessment. Lastly, the results underscore the immense value of ocular data; the inclusion of blink markers consistently provides the largest incremental boost in classification accuracy, confirming that blink dynamics offer complementary, discriminative information beyond cardio-respiratory signals for assessing cognitive workload.

6.5 RGB Algorithmic Baseline for rPPG

In Tables 6 & 7 we report the baseline algorithmic methods for the RGB modality. We notice model-driven approaches such as POS [45] and CHROM [10] significantly outperform simpler method such as Green [43] and ICA [36]. This can be attributed to the inductive biases within the algorithms to denoise and mitigate the effects of motion. Green is a simple single-channel averaging operation, while ICA assumes independence with noise and the rPPG signals. However, these methods work well under simple, stationary cases. In the presence of motion, and thereby, sub-surface scattering changes due to motion, these assumptions are no longer true. POS and CHROM are projection methods derived from optical and physiological properties of the skin and blood, leading them to outperform the other baselines.

6.6 Bias Analysis for rPPG

Bias in rPPG is has been extensively documented [7, 44, 34, 9]. The rPPG signal is the diffuse component of light that reflects off the skin. That is, it captures pulsatile changes through sub-surface scattering. However, light rays are attenuated as they travel through the epidermal layers. Melanin is one such reason. High concentration of melanin causes more light to be absorbed, thereby reducing the signal strength of the incident light ray. In other words, the SNR of the observed rPPG signals is weaker for participants/volunteers with darker skin tones.

In accordance with this, we report the performance of the rPPG algorithms across several skin tones in Tables 10 & 11. Specifically, we categorize the skin tone into *light*, *medium*, and *dark* by binning

Table 10: **Performance of standard rPPG algorithms across skin tones - HR Metrics.** The metrics for *light*, *medium*, and *dark* skin tones have been reported together, spaced by a forward slash. 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.

		HR Metrics (BPM) **Data Split: *light* / *medium* / *dark***			
	Method	MAE ↓	RMSE ↓	MAPE ↓	r ↑
NIR	**PhysNet** [50]	13.39 / 13.75 / 13.12	16.67 / 17.16 / 16.18	17.19 / 18.14 / 15.32	-0.11 / -0.04 / -0.01
	RythmFormer [55]	4.90 / 9.12 / 9.03	7.88 / 11.93 / 11.84	5.70 / 11.46 / 10.14	0.68 / 0.53 / -0.19
	PhysFormer [52]	11.46 / 12.02 / 11.07	14.75 / 15.42 / 12.91	15.09 / 16.33 / 13.54	-0.02 / -0.01 / 0.09
	FactorizePhys [18]	8.50 / 9.42 / 10.83	15.02 / 15.98 / 17.48	9.78 / 11.16 / 12.44	0.32 / 0.47 / 0.25
	PhysMamba [27]	4.90 / 6.38 / 5.22	9.48 / 11.77 / 8.75	6.21 / 9.28 / 6.19	0.64 / 0.55 / 0.65
	Contrast-Phys+ [41]	6.75 / 8.54 / 6.14	11.13 / 13.73 / 9.60	8.42 / 11.98 / 7.08	0.48 / 0.30 / 0.58
	Pretrained Contrast-Phys+	4.36 / 6.71 / 5.20	8.62 / 12.35 / 8.41	5.28 / 9.70 / 6.09	0.71 / 0.47 / 0.66
RGB	**Green** [43]	22.42 / 22.56 / 24.20	27.33 / 27.45 / 28.65	26.84 / 26.97 / 28.27	-0.15 / 0.07 / -0.03
	ICA [36]	13.70 / 13.30 / 16.72	19.83 / 20.22 / 21.97	16.12 / 15.42 / 19.29	0.21 / 0.24 / 0.05
	CHROM [10]	3.74 / 5.86 / 8.18	7.60 / 10.71 / 12.56	4.48 / 7.01 / 9.37	0.79 / 0.69 / 0.39
	POS [45]	3.10 / 6.01 / 7.14	6.51 / 13.90 / 11.10	3.88 / 7.42 / 8.33	0.84 / 0.59 / 0.47
	PhysNet [50]	11.70 / 11.45 / 10.01	15.64 / 15.40 / 13.43	14.83 / 15.16 / 11.67	0.00 / 0.14 / 0.22
	RythmFormer [55]	4.11 / 10.16 / 7.62	6.55 / 12.72 / 10.46	4.89 / 12.87 / 8.44	0.75 / 0.43 / 0.04
	PhysFormer [52]	8.85 / 10.46 / 9.42	14.31 / 15.08 / 12.57	11.14 / 14.10 / 11.46	0.14 / 0.16 / 0.01
	FactorizePhys [18]	3.52 / 4.14 / 7.69	9.04 / 9.72 / 13.85	4.09 / 4.81 / 8.65	0.74 / 0.76 / 0.38
	PhysMamba [27]	3.53 / 4.35 / 5.44	8.90 / 9.18 / 8.71	4.04 / 5.55 / 6.55	0.72 / 0.75 / 0.61
	Contrast-Phys+ [41]	2.93 / 5.00 / 5.23	6.20 / 10.32 / 8.72	3.42 / 7.27 / 5.98	0.86 / 0.65 / 0.63
	Fusion	3.07 / 4.90 / 4.98	7.39 / 10.23 / 8.51	3.55 / 7.32 / 5.68	0.80 / 0.66 / 0.66

Table 11: **Performance of standard rPPG algorithms across skin tones - HRV Metric.** The metrics for *light*, *medium*, and *dark* skin tones have been reported together, spaced by a forward slash. We report SNR, MACC, AAMI, and IBI error for HRV estimation.

		Data Split: *light* / *medium* / *dark*			
	Method	SNR (dB) ↑	MACC ↑	AAMI (%) ↑	IBI (ms) ↓
NIR	**PhysNet** [50]	-6.09 / -5.96 / -5.90	0.16 / 0.18 / 0.19	25.00 / 22.97 / 23.33	327.25 / 323.71 / 247.68
	RythmFormer [55]	2.08 / -4.42 / -2.12	0.62 / 0.44 / 0.48	67.39 / 35.83 / 41.67	48.41 / 88.85 / 73.33
	PhysFormer [52]	-7.66 / -8.10 / -8.97	0.30 / 0.31 / 0.31	28.81 / 28.05 / 21.67	115.53 / 120.57 / 99.22
	FactorizePhys [18]	-1.90 / -3.05 / -3.71	0.52 / 0.50 / 0.47	63.98 / 59.76 / 57.50	82.48 / 85.01 / 75.04
	PhysMamba [27]	-0.04 / -1.38 / -1.82	0.58 / 0.54 / 0.52	74.58 / 68.90 / 70.83	68.32 / 80.50 / 53.05
	Contrast-Phys+ [41]	-0.88 / -2.19 / -1.84	0.55 / 0.50 / 0.52	62.71 / 55.69 / 60.00	62.55 / 83.70 / 54.13
	Pretrained Contrast-Phys+	0.73 / -0.84 / -0.89	0.62 / 0.57 / 0.57	76.69 / 66.26 / 67.50	56.72 / 78.37 / 51.93
RGB	**Green** [43]	-8.17 / -8.26 / -9.01	0.24 / 0.24 / 0.22	25.42 / 18.95 / 13.33	125.64 / 117.80 / 121.75
	ICA [36]	-3.57 / -3.80 / -6.18	0.38 / 0.39 / 0.32	48.31 / 47.78 / 32.50	69.86 / 71.19 / 93.43
	CHROM [10]	-1.52 / -2.22 / -4.21	0.44 / 0.42 / 0.37	82.20 / 70.56 / 57.50	41.88 / 63.64 / 69.98
	POS [45]	-1.11 / -1.46 / -2.30	0.44 / 0.44 / 0.42	84.32 / 74.18 / 62.50	44.67 / 54.10 / 52.56
	PhysNet [50]	-4.79 / -4.84 / -4.90	0.20 / 0.21 / 0.20	36.86 / 37.10 / 35.00	282.20 / 286.58 / 245.46
	RythmFormer [55]	3.15 / -4.42 / -2.46	0.50 / 0.37 / 0.43	76.09 / 30.83 / 45.83	41.69 / 98.16 / 57.33
	PhysFormer [52]	-0.89 / -3.73 / -4.61	0.40 / 0.37 / 0.38	51.27 / 42.14 / 43.33	86.10 / 102.85 / 76.70
	FactorizePhys [18]	1.00 / 0.22 / -2.33	0.49 / 0.48 / 0.43	88.98 / 82.46 / 68.33	36.51 / 34.74 / 65.09
	PhysMamba [27]	1.39 / 0.49 / -1.02	0.49 / 0.48 / 0.45	88.14 / 80.24 / 70.00	36.43 / 54.29 / 49.05
	Contrast-Phys+ [41]	2.15 / 1.23 / -0.54	0.51 / 0.50 / 0.47	87.29 / 78.43 / 72.50	33.69 / 55.64 / 51.79
	Fusion	2.42 / 1.36 / -0.13	0.52 / 0.50 / 0.47	89.83 / 78.66 / 73.33	34.35 / 54.56 / 47.47

the 6 Fitzpatrick skin-tone labels (two per bin) [38]. This binning ameliorates noise in label collection. Participants with lighter skin tones show the best performance across all modalities, reflecting a fundamental challenge in optical physiological sensing: melanin in darker skin absorbs more light in the visible spectrum, making it harder to detect the subtle blood volume changes needed for HR measurement. NIR wavelengths experience less melanin absorption, which is why NIR methods generally show smaller performance drops across skin tones—the physics favors more equitable sensing. RGB achieves better absolute accuracy on participants with lighter skin tones but suffers steeper degradation on participants with darker skin tones due to this melanin interference. Fusion delivers the best overall performance by combining both modalities, achieving strong accuracy on participants with lighter skin tones while maintaining relative stability on participants with medium and darker skin tones—essentially leveraging RGB's peak performance where it works well and NIR's robustness where RGB struggles. However, Fusion doesn't eliminate the bias; it still shows meaningful degradation from participants with lighter to medium skin tones before stabilizing. The

fairness advantage of NIR stems from its reduced melanin sensitivity, but even NIR methods show performance gaps, indicating that skin tone bias in remote sensing is partly a hardware problem rooted in optical physics, not just algorithmic design. Further, we note that the distribution of ski tone labels is not uniform - 10, 21, and 5 for participants with lighter, medium, and darker skin tone - and can introduce uncertainty in the estimates.

7 Cognitive Load Prediction Benchmarking Results

7.1 Detailed Performance Analysis

The evaluation across seven experimental configurations revealed several key insights about physiological signal utility for cognitive load detection. Note: The metrics numbers in this section are from the test split in the main paper, not the 4-fold cross-validation in 6.4.

Contact vs. Remote Sensing Performance Contact PPG achieved 70.00% accuracy with Random Forest, establishing a strong baseline for cardiac-based cognitive load detection. rPPG performance was notably lower at 56.41%, reflecting the inherent noise and artifacts in camera-based physiological sensing. However, this gap was substantially reduced through multimodal integration, with remote sensing combinations ultimately achieving superior performance to any single contact modality.

Multimodal Integration Benefits The systematic addition of modalities demonstrated clear performance improvements (models are the best-performing ones of that modality combination):

- rPPG alone: 56.41% (RF)
- rPPG + Remote Respiratory: 73.08%(RF)
- rPPG + Remote Respiratory + Blink: 86.49% (GB)

This progression illustrates the complementary nature of different physiological indicators, with each modality contributing unique information about cognitive state.

Model Architecture Comparison Traditional ML models consistently outperformed DL approaches across most configurations. The best-performing models were:

- Gradient Boosting: Highest overall performance (86.49% accuracy)
- Random Forest: Consistent strong performance across configurations
- SVM: Competitive performance with good generalization
- DL: Best performance with blink-only data (68.00% CNN)

Feature Importance Insights Analysis of feature importance in the best-performing models revealed that HRV metrics (RMSSD, SDNN, pNN50) and blink dynamics (rate, duration variability) were the most discriminative features for cognitive load classification. This aligns with physiological literature on autonomic nervous system responses to mental workload.

Generalization Performance The participant-based split strategy ensured that all reported results represent true generalization to unseen individuals. The consistent performance across different participants validates the robustness of the extracted physiological features and the potential for deployment in real-world scenarios without requiring user-specific calibration.

7.2 PPG Feature Importance Analysis

To understand which PPG-derived features were most influential in the classification decisions, after training the cognitive load classifier with contact-PPG features only, we averaged the feature importances across all 37 Leave-One-Subject-Out folds for Random Forest and XGBoost models. The top 10 most important features for each model are listed in Table 12.

Both models consistently ranked features related to HR and HRV as highly important. Mean BPM and the average IBI were top contributors for both the Random Forest (RF) and Gradient Boosting (GB)

Table 12: **Top 10 Mean Feature Importances from LOSO CV.** PPG features are described in detail in 5.1.1.

Random Forest			Gradient Boosting		
Rank	**Feature**	**Mean Importance**	**Rank**	**Feature**	**Mean Importance**
1	bpm	0.107	1	bpm	0.122
2	ibi	0.103	2	sd1/sd2	0.119
3	sdnn	0.095	3	pnn50	0.117
4	pnn20	0.090	4	pnn20	0.115
5	sd1/sd2	0.090	5	sdnn	0.110
6	pnn50	0.088	6	ibi	0.110
7	sd2	0.087	7	sd2	0.068
8	s	0.077	8	sdsd	0.062
9	sdsd	0.070	9	s	0.045
10	sd1	0.066	10	rmssd	0.038

classifiers, aligning with the expectation that overall HR changes under cognitive load. Specifically, SDNN (representing overall HRV), the Proportion of successive NN intervals differences pNN20 and pNN50 (associated with parasympathetic nervous system activity), and the Poincaré plot ratio of the standard deviation (SD1) to that along the line-of-identity (SD2), reflecting short-term to long-term variability, appeared important for both classifiers. The high ranking of these features is consistent with physiological literature suggesting that increased cognitive load often leads to increased HR (i.e., lower IBI) and decreased HRV (i.e., lower SDNN, pNN50, RMSSD, and potentially altered Poincaré metrics).

7.3 Subject-Level Performance Analysis

To better understand the generalization patterns of our cognitive load classification models, we conducted a detailed subject-level performance analysis using the best-performing configuration (rPPG + Remote Respiratory + Blink Markers) with the Gradient Boosting classifier. This analysis provides insights into individual differences in physiological responses to cognitive load and the corresponding classification performance.

Experimental Methodology For this analysis, we extracted features from all three modalities (rPPG, remote respiratory, and blink markers) for each subject in the test set (n=10) and trained the Gradient Boosting model on the training set with the same hyperparameters used in the main experiments. We then evaluated performance separately for each subject, calculating per-subject accuracy, F1 score, specificity, and sensitivity. Additionally, we analyzed performance variations across different cognitive tasks to identify which tasks were most consistently classified correctly.

Subject-Level Performance Distribution The per-subject analysis revealed substantial variability in classification performance. As shown in Figure 8, 7 out of 10 test subjects achieved perfect classification accuracy (100%), demonstrating robust performance across the majority of participants. However, three subjects (v22, v32, and v23) showed notably lower performance, with accuracies of 75%, 50%, and 33.3%, respectively.

Analysis of Difficult Cases Examining the confusion matrices for the lowest-performing subjects revealed interesting patterns. Subject v23 showed consistent misclassification of high cognitive load states as low load (false negatives), suggesting this individual may have exhibited physiological responses that deviated from the typical patterns observed in the training population. In contrast, subject v32 exhibited more balanced errors, with both false positives and false negatives, indicating a general difficulty in distinguishing cognitive load states for this participant based on the extracted physiological features.

Task-Specific Performance Analysis across different task types revealed that number tasks (arithmetic) were most consistently classified correctly (90% accuracy), while reading tasks achieved 77.8% accuracy. This suggests that arithmetic tasks may elicit more consistent and distinctive physiological

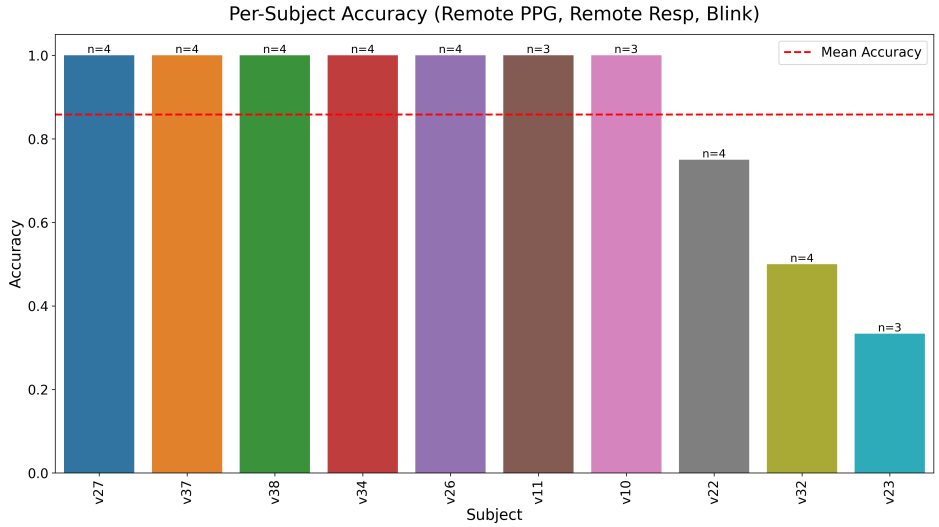


Figure 8: Per-subject classification accuracy for cognitive load detection using the Gradient Boosting model with rPPG + Remote Respiratory + Blink features. Annotations indicate the number of samples per subject, with most subjects having 4 samples. Seven subjects achieved perfect classification, while three subjects (v22, v32, and v23) showed substantially lower performance.

responses across subjects compared to reading tasks. The still (baseline) task showed moderate classification performance, likely due to individual variations in resting physiological patterns.

7.4 Cognitive Load Cross-Validation and Statistical Analysis

To ensure robust performance estimates, we conducted 4-fold cross-validation (fold split detailed in section 6.4) for cognitive load classification using the best-performing configuration (rPPG + Remote Resp + Blink Markers).

Cross-Validation Results: Gradient Boosting achieved the best performance across folds:

- Mean Accuracy: 0.832 ± 0.012 (95% CI: $[0.813, 0.851]$)
- Mean F1 Score: 0.855 ± 0.015
- Mean Precision: 0.814 ± 0.042
- Mean Recall: 0.907 ± 0.068

Statistical Significance Testing: We conducted comprehensive statistical analyses across 8 ML models with 4-fold cross-validation (32 total evaluations):

- *One-way ANOVA:* Revealed significant differences between models in accuracy ($F = 12.55$, $p = 0.00011$), rejecting the null hypothesis that all models perform equivalently.
- *Pairwise t-tests:* With Bonferroni correction ($\alpha = 0.0018$), SVM significantly outperforms multiple baseline models (e.g., vs. Decision Tree, $p < 0.01$).
- *Tukey HSD post-hoc tests:* Confirmed significant performance differences for Gradient Boosting compared to several baseline models.
- *Coefficient of Variation:* SVM exhibits low variability (CV < 0.02 for accuracy and F1) across folds, which shows stability in performance.

This statistical validation confirms that Gradient Boosting with multimodal remote signals provides significantly better and more consistent cognitive load classification compared to alternative approaches.

7.5 Cognitive Load Classification Discussions

We evaluated ML benchmark models for cognitive load classification using the physiological features extracted from various signal combinations. The results demonstrate the effectiveness of multimodal signal integration for cognitive load assessment.

For the result metrics, ML models with crafted, physiology-relevant features consistently outperformed DL approaches across most signal configurations. The Gradient Boosting classifier achieved the highest overall accuracy of 86.49% (F1: 0.878) when utilizing the complete multimodal signal set (rPPG + Remote Respiratory + Blink Markers). This represents a substantial improvement over unimodal approaches, with rPPG alone achieving 69.23% and Blink Markers alone reaching 65.00%.

Across all ML classifiers, the integration of multiple physiological signals consistently yielded superior performance compared to single-signal approaches. Notably, the addition of blink markers to cardiorespiratory signals produced the largest performance gains, with accuracy improvements ranging from $8 - 18\%$ over the next best configuration. This suggests that ocular dynamics contain complementary information to cardiac and respiratory patterns when assessing cognitive workload.

For the DL models, a different pattern emerged. The 1D CNN and ResNet1D architectures performed best with blink markers alone (68.00%accuracy), while their performance on multimodal inputs was notably inferior to traditional ML approaches. This suggests that the selected deep architectures may be less effective at integrating the heterogeneous information present in multimodal physiological signals compared to ensemble methods like Gradient Boosting and Random Forest. Our CogPhys dataset presents an opportunity for future research to develop specialized DL architectures specifically designed for multimodal physiological signal integration, potentially through attention mechanisms or specialized fusion strategies that better capture cross-modal interactions.

While both contact-based and rPPG signals alone showed moderate classificatory power, neither achieved particularly strong results in isolation (70.00% and 56.41% with RF, respectively). However, this performance limitation was dramatically overcome through multimodal integration, with the combination of remote sensing modalities ultimately achieving the highest overall performance. This finding underscores the importance of complementary signal sources in remote physiological sensing applications, demonstrating that the integration of multiple less-than-ideal remote signals can effectively surpass the performance of individual high-fidelity contact measurements.

When comparing across model types, ensemble methods (particularly Gradient Boosting) consistently outperformed other approaches, likely due to their robustness to the high variability and non-linear relationships present in physiological data, as well as their ability to effectively utilize the carefully crafted, physiologically meaningful features extracted from each modality. Our participant-based test split, where models were evaluated on previously unseen subjects, ensures these results represent generalizable performance across individuals, validating the potential for practical deployment in real-world settings where new users would not require personalized calibration.

8 Ethical Safeguards and Data Privacy

8.1 Data Privacy and Security Measures

Our dataset was collected under IRB approval (Rice University IRB-FY2025-59) with informed consent from all participants. We implement multiple layers of data protection:

Secure Storage: All dataset files are stored on Rice University's Box Drive, an IRB-approved secure cloud storage solution with enterprise-grade encryption and access controls.

Access Control: Dataset access is regulated through a comprehensive Data Use Agreement (DUA). Researchers requesting access must:

- Provide institutional affiliation verification
- Agree to use data solely for non-commercial research purposes
- Commit to proper anonymization practices when using data in publications
- Acknowledge time-gated access periods to limit misuse from extended link durations

Anonymization Guidelines: For participants who consented to having their data used in publications and presentations, we provide clear guidelines on face anonymization techniques (blurring, masking) when required. The dataset folder is accompanied by a list of participants whose faces must not be used in illustrative formats such as publications, presentations, websites, and other similar formats. The test set for the main fold was specifically constructed to include only participants who provided such consent, enabling researchers to use these samples for presentations. The researchers, however, must still adhere to anonymization protocols such as blurring the face and masking the eyes.

8.2 Long-term Societal and Ethical Implications

Remote cognitive load monitoring presents significant opportunities and challenges for society:

Positive Applications:

- Enhanced user experience and convenience for drivers.
- Improved accessibility of cognitive assessment in telehealth settings
- Development of adaptive user interfaces that respond to mental workload
- Support for assistive technologies for individuals with cognitive impairments

Ethical Concerns:

- *Privacy:* Continuous camera-based physiological monitoring raises concerns about consent and data collection in public/private spaces
- *Algorithmic Bias:* Performance variations across demographic groups (see Section 6.6) require careful consideration to ensure equitable outcomes
- *Workplace:* Potential misuse for employee monitoring without proper consent or oversight
- *Data Security:* Physiological data represents sensitive biometric information requiring robust protection

Recommendations: We advocate for responsible development that includes:

- Clear disclosure and opt-in consent for any cognitive monitoring applications
- Regular algorithmic audits for bias across diverse populations
- Strict data minimization principles (collecting only necessary data)
- Transparent reporting of system limitations and failure modes
- Multi-stakeholder engagement in deployment decisions

9 Implications, Broader Impacts, and Limitations

The evaluation of cognitive load classification using multimodal physiological sensing reveals several important implications for the field of human-computer interaction and cognitive state monitoring:

Practical Deployment Considerations The superior performance of traditional ML models with engineered features suggests that domain knowledge remains crucial for physiological signal analysis. While DL approaches offer the promise of end-to-end learning, the complexity and variability of physiological signals may require more sophisticated architectures or larger datasets to achieve comparable performance. The 86.49% accuracy achieved with multimodal remote sensing approaches the performance levels needed for practical applications in cognitive load monitoring systems.

Multimodal Sensing Advantages The substantial performance gains from multimodal integration (from 56.41% with rPPG alone to 86.49% with the full multimodal setup) demonstrate the critical importance of complementary physiological indicators. This finding has significant implications for the design of future cognitive monitoring systems, suggesting that investment in multiple sensing modalities yields substantial returns in classification accuracy.

Remote Sensing Viability The ability to achieve high classification accuracy using entirely remote sensing modalities (camera-based rPPG, thermal respiratory sensing, and computer vision-based blink detection) validates the feasibility of non-intrusive cognitive load monitoring. This capability is particularly valuable for applications where contact sensors are impractical or undesirable, such as long-term monitoring, automotive applications, or public spaces.

Broader Impacts Remote cognitive load monitoring has significant societal implications. Positive impacts can include improved accessibility of cognitive assessment in stationary healthcare settings, and drivinassistiveve technologies. However, these benefits come with important ethical considerations, including (1) privacy concerns with continuous camera-visual or physiological monitoring, (2) potential algorithmic bias across diverse populations, and (3) risks of workplace surveillance. We advocate for responsible development that balances innovation with appropriate safeguards for privacy, equity, and individual autonomy.

Limitations and Future Directions Several limitations should be acknowledged: (1) The binary classification approach (Low vs. High cognitive load) may not capture the full spectrum of cognitive states; (2) The participant-based split, while ensuring generalization, resulted in a relatively small test set (10 participants); (3) The laboratory setting may not fully represent real-world conditions with additional confounding factors; (4) The 120-second signal windows may not be optimal for all applications requiring real-time monitoring.

Generalizability Considerations While the participant-based evaluation strategy ensures that results represent true generalization to unseen individuals, the demographic composition of our dataset (primarily young adults, university setting) may limit generalizability to broader populations. Future work should evaluate performance across diverse age groups, occupations, and cultural backgrounds to establish the robustness of these approaches for widespread deployment.

References

[1] Sarah Allred, Sean Duffy, and John Smith. Cognitive load and strategic sophistication. *Journal of Economic Behavior & Organization*, 125:162–178, 2016.

[2] Prithila Angkan, Behnam Behinaein, Zunayed Mahmud, Anubhav Bhatti, Dirk Rodenburg, Paul Hungler, and Ali Etemad. Multimodal brain–computer interface for in-vehicle driver cognitive load measurement: Dataset and baselines. *IEEE Transactions on Intelligent Transportation Systems*, 25(6):5949–5964, 2024.

[3] 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.

[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] Anubhav Bhatti, Prithila Angkan, Behnam Behinaein, Zunayed Mahmud, Dirk Rodenburg, Heather Braund, P James Mclellan, Aaron Ruberto, Geoffery Harrison, Daryl Wilson, et al. Clare: Cognitive load assessment in realtime with multimodal data. *arXiv preprint arXiv:2404.17098*, 2024.

[6] 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.

[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] Ananyananda Dasari, Sakthi Kumar Arul Prakash, László A Jeni, and Conrad S Tucker. Evaluation of biases in remote photoplethysmography methods. *NPJ digital medicine*, 4(1):91, 2021.

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

[11] Cary Deck and Salar Jahedi. The effect of cognitive load on economic decision making: A survey and new experiments. *European Economic Review*, 78:97–119, 2015.

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

[13] 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.

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

[15] Martin Gjoreski, Tine Kolenik, Timotej Knez, Mitja Luštrek, Matjaž Gams, Hristijan Gjoreski, and Veljko Pejović. Datasets for cognitive load inference using wearable sensors and psychological traits. *Applied Sciences*, 10(11):3843, 2020.

[16] 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.

[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] Apostolos Kalatzis, Ashish Teotia, Vishnunarayan Girishan Prabhu, and Laura Stanley. A database for cognitive workload classification using electrocardiogram and respiration signal. In *International Conference on Applied Human Factors and Ergonomics*, pages 509–516. Springer, 2021.

[21] Thomas Kosch, Jakob Karolus, Johannes Zagermann, Harald Reiterer, Albrecht Schmidt, and Paweł W Woźniak. A survey on measuring cognitive workload in human-computer interaction. *ACM Computing Surveys*, 55(13s):1–39, 2023.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] Valentina Markova, Todor Ganchev, and Kalin Kalinkov. Clas: A database for cognitive load, affect and stress recognition. In *International Conference on Biomedical Innovations and Applications*, pages 1–4. IEEE, 2019.

[29] 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.

[30] Igor Mijić, Marko Šarlija, and Davor Petrinović. Mmod-cog: A database for multimodal cognitive load classification. In *11th International Symposium on Image and Signal Processing and Analysis*, pages 15–20. IEEE, 2019.

[31] Wim De Neys. Dual processing in reasoning: Two systems but one reasoner. *Psychological science*, 17(5):428–433, 2006.

[32] 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.

[33] 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.

[34] Ewa M Nowara, Daniel McDuff, and Ashok Veeraraghavan. A meta-analysis of the impact of skin tone and gender on non-contact photoplethysmography measurements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 284–285, 2020.

[35] 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.

[36] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1):7–11, 2010.

[37] Andrea Révész, Marije Michel, and Roger Gilabert. Measuring cognitive task demands using dual-task methodology, subjective self-ratings, and expert judgments: A validation study. *Studies in second language acquisition*, 38(4):703–737, 2016.

[38] Silonie Sachdeva. Fitzpatrick skin typing: Applications in dermatology. *Indian journal of dermatology, venereology and leprology*, 75:93, 2009.

[39] Stefan Schneegass, Bastian Pfleging, Nora Broy, Frederik Heinrich, and Albrecht Schmidt. A data set of real world driving to assess driver workload. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 150–157, 2013.

[40] 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.

[41] 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.

[42] Paul Van Gent, Haneen Farah, Nicole Van Nes, and Bart Van Arem. Heartpy: A novel heart rate algorithm for the analysis of noisy signals. *Transportation research part F: traffic psychology and behaviour*, 66:368–378, 2019.

[43] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] Gilsang Yoo, Hyeoncheol Kim, and Sungdae Hong. Prediction of cognitive load from electroencephalography signals using long short-term memory network. *Bioengineering*, 10(3):361, 2023.

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

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

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