Classification of Infant Sleep–Wake States from Natural Overnight In-Crib Sleep Videos

Shayda Moezzi[1], Michael Wan[1], Sai Kumar Reddy Manne[1], Amal Mathew[1], Shaotong Zhu[1], Bishoy Galoaa[1], Elaheh Hatamimajoumerd[1], Emma C. Grace[2], Cassandra B. Rowan[2], Emily Zimmerman[1], Briana J. Taylor[1], Marie J. Hayes[2], Sarah Ostadabbas[1*]
[1]Northeastern University, Boston, MA, USA
[2]University of Maine, Orono, ME, USA
[*]Corresponding author: ostadabbas@ece.neu.edu

Abstract

Infant sleep is critical for healthy development, and disruptions in sleep patterns can have profound implications for infant brain maturation and overall well-being. Traditional methods for monitoring infant sleep often rely on intrusive equipment or time-intensive manual annotations, which hinder their scalability in clinical and research applications. We present our dataset, SmallSleeps, which includes 152 hours of overnight recordings of 17 infants aged 4–11 months captured in real-world home environments. Using this dataset, we train a deep learning algorithm for classification of infant sleep–wake states from short 90 s video clips drawn from natural, overnight, in-crib baby monitor footage, based on a two-stream spatiotemporal model which integrates rich RGB frames with optical flow features. Our binary classification algorithm was trained and tested on "pure" state clips featuring a single state dominating the timeline (i.e., over 90% sleep or over 90% wake) and achieves over 80% precision and recall. We also perform a careful experimental study of the result of training and testing on "mixed" clips featuring specified levels of heterogeneity, with a view towards applications to infant sleep segmentation and sleep quality classification in longer, overnight videos, where local behavior is often mixed. This local-to-global approach allows for deep learning to be effectively deployed on the strength of tens of thousands of video clips, despite a relatively modest sample size of 17 infants[1].

1. Introduction

Sleep plays a critical role in the physiological and neurological development of infants during their early life [19]. Infant sleep patterns undergo significant changes, typically transitioning within the first few months of life from an ultradian rhythm with multiple hours-long sleep cycles per day to a circadian rhythm aligned with a 24-hour cycle. Recent research has demonstrated a strong connection between neonatal development and sleep cycle regularity in infants [1, 8].

The gold standard for sleep monitoring remains polysomnography (PSG) [23], which involves recording various physiological signals over a night using multiple contact sensors. These signals are then manually analyzed by a sleep technician to annotate sleep stages as defined by the American Academy of Sleep Medicine [14]. However, PSG presents several challenges for continuous sleep monitoring. It typically requires the subject to stay overnight in a dedicated sleep lab, supervised by a technician, with multiple sensors attached to the body for extended periods. This setup can cause discomfort and disrupt natural sleep cycles, particularly problematic for newborns with sensitive skin that is prone to irritation from contact-based sensors [2].

The invasive and costly nature of PSG underscores the need for alternative, non-contact sleep monitoring techniques, particularly for neonates and other sensitive populations. Wearable devices, such as wrist or ankle monitors, have been explored to ease participant burden, but still require direct contact [26, 28]. Increasingly, video-based solutions have emerged as a viable option for non-invasive sleep monitoring. These "videosomnography" methods rely on recordings from video cameras, which can be easily installed in homes and healthcare facilities without causing discomfort to the subject [29]. Such approaches range from manual expert behavioral coding of video recordings to automated sleep scoring using machine learning and computer

[1]Our code can be found at https://github.com/ostadabbas/Infant-Sleep-vs-Awake-Detection. Supported by NSF-CAREER Grant #2143882, a Northeastern University–University of Maine Seed Grant, and a Northeastern University TIER 1 Seed Grant.

vision algorithms [21].

Videosomnography is particularly well-suited for infant sleep studies, as many parents are already accustomed to using video monitors for overnight observation. Despite the rapid advancements in deep learning and computer vision for infant pose and state estimation [13,27,31], current video-based pediatric sleep monitoring approaches are often limited. Many rely on basic motion feature analysis using simplistic machine learning algorithms or are confined to highly controlled environments, such as neonatal intensive care units (NICUs). Some attempts to make use of deep learning have been limited to qualitative studies. A detailed review of these prior methods is provided in Section 2.

1.1. Our Contributions

In this paper, we advance the field of pediatric videosomnography by introducing a novel spatiotemporal network for classifying sleep and wake states from video footage. This work is grounded in our annotated dataset, SmallSleeps, comprised of 152 hours of overnight video recordings from 17 infant subjects, with video-based behavioral coding for sleep–wake states. The rich set of annotations makes our dataset one of the most comprehensive of its kind, despite the relatively small sample size.

We focus specifically on the task of video classification of sleep–wake states in short 90 s clips extracted from our overnight video dataset. Each clip is assigned an overall label based on the majority state present in the clip, as determined by the behavioral coding. Recognizing the complexity of mixed-state clips—those containing transitions between sleep and wake—we explore the impact of training set enrichment by selectively including clips with higher purity (e.g., $\geq$90% sleep or $\geq$90% wake), which will inform future efforts in dataset curation and algorithm design. We also experimentally quantify performance results on test sets with varying states of purity, giving a nuanced and realistic picture of the challenges to come in infant sleep segmentation over long videos.

Architecturally, we explore the use of two-stream neural networks which can adaptively fuse spatiotemporal signals extracted from both the RGB video frames and derived optical flow. This ability to draw selectively from the semantically-rich RGB frames and the motion-attune optical flow frames has been found to be effective for the detection of subtle, short term infant behavioral signals in the crib, such as pacifier sucking [31]. Our work here shows that this approach can also be effective at discerning more complex, holistic behavioral state changes, namely, between sleep and arousal periods reflecting state stability.

1.2. Small Data Statement

Classifying infant sleep–wake states from video data is challenging due to the difficulty of acquiring data given unique privacy and logistical considerations and the inherent behavioral complexity of distinguishing infant sleep states. The visual diversity in infant videos—including differences in clothing, body coverage, and camera angles—further complicates model training and necessitates robust generalization capabilities from algorithms trained on these limited datasets.

This study qualifies as small data research because of the scarcity of annotated infant sleep video data due to these constraints. To overcome this challenge, we leverage state-of-the-art models pre-trained on large-scale datasets, incorporating both RGB and spatiotemporal features to build an end-to-end pipeline optimized for overnight infant video analysis. This approach allows us to identify a model capable of generalizing effectively across diverse conditions, addressing the inherent challenges of small and heterogeneous datasets while advancing infant sleep state detection.

2. Related Work

Contact-free videosomnography has emerged as an attractive alternative to polysomnography for sleep studies, given its potential to significantly reduce costs and both clinician and patient burden, and increase accessibility to sleep diagnoses and medicine. Wearable trackers address many of these issues, but the scope of what they can detect is limited, and they are still invasive, especially in the pediatric domain, where they might harm the sensitive skin of infants and interfere with sleep patterns. Here, we review prior research in automatic video-based infant sleep coding, mostly relying on classical machine learning and computer vision. These approaches are summarized in Table 1.

2.1. Early Approaches to Automatic Infant Videosomnography

In the late 2010s, a first generation of algorithmic infant videosomnography emerged, with researchers combining classical computer vision and machine learning tools, often extracting global motion features or detecting motion actigraphy visually. Pioneering work by Schwichtenberg et al. from 2018 [21] isolated movement from the image foreground and fed those parameters into the Sadeh actigraphy algorithm [20] based on linear discriminant analysis. A relatively large population of 30 infants and toddlers was used for testing, but the algorithm suffered from low specificity. Cabon et al. [5] published a study in 2019 using machine learning classifiers based on features from audio processing and computer vision motion and eye tracking, achieving high accuracy for detection of some alertness states for ten newborns in the NICU. However, their eye tracking method relies on human guidance, and the detection concordance across all sleep stages is only moderate. Barnett et al. [4] published an abstract in 2019 claiming good results from computer vision-based sleep–wake classification on short

Table 1. Infant focused prior videosomnography studies.

Work	N	Ages (mo.)	Setting	Task	Method
Schwichtenberg et al. (2018) [21]	30	8–30	Home	Sleep–wake segmentation	Movement-based foreground motion index, Sadeh algorithm [20]
Cabon et al. (2019) [5]	10	Newborns	NICU	Sleep stage classification	ML classification from vocalization, CV-based motion, and semi-manually-guided eye tracking features
Barnett et al. (2019, abstract) [4]	7	0–24	Clinic	Sleep–wake classification	Proprietary "computerized procedure of machine learning"
Long et al. (2019) [17]	10	3–9	Lab	Sleep–wake classification	CV actigraphy, Bayesian linear discriminant classifier
Mukai et al. (2019) [18]	8	Newborns	NICU	6-way sleep–wake state classification	Face detection & alignment preprocessing, ML classifier on spatiotemporal features
Awais et al. (2021) [3]	21	Newborns	NICU	Sleep–wake classification	Color-based face crop, frame-based CNN features, ML classifier
Khan et al. (2021) [16]	N/A	Dolls	N/A	Qualitative analysis	Hypothetical sleep–wake detection based on eye detection
Choi et al. (2023) [7]	729	All	Hospital	Sleep apnea segmentation	(2+1)D spatiotemporal CNN classifier
Singh et al. (2023) [22]	0	N/A	N/A	Sleep–wake classification	Hypothetical rule-based system from pose and eye detection
Huang et al. (2024) [12]	103	Preterm & term	NICU	Sleep–wake classification	Frame-based CNN features with fully connected classifier
Ours (2024)	17	4–11	Home	Sleep–wake classification	2-stream fusion of 3D convolution with self-attention

N: number of subjects, ML: machine learning, DL: deep learning, CV: computer vision, CNN: convolutional neural network.

video clips drawn from seven infants, from the team's Nanit commercial baby monitor, but few details are disclosed. In 2019, Long et al. [17] published a well-regarded study using computer vision actigraphy features for Bayesian linear discriminant classification of sleep–wake, achieving 0.73 Cohen's κ concordance with PSG, on a population of 10 infants in 1 h videos. In the same year, Mukai et al. [18] presented conference results using face detection and spatiotemporal features with machine learning classifiers, but struggled with their 6-way classification problem, in part due to class imbalance.

2.2. Recent Deep Learning Techniques

The 2020s have seen the introduction of early deep learning approaches. In 2021, Awais et al. [3] published a comprehensive study using color-based face cropping, frame-based convolutional features, and a support vector machine classifier, achieving a 93% F1 score on 2 h videos from 19 newborns in the NICU. However, their method relies on clean face detections, likely only possible due to their highly controlled NICU setting. Moreover, their approach frame-based, rather than spatiotemporal, making it unlikely that their algorithm can incorporate infant movement information required for sleep–wake classification in less pristine settings. Also in 2021, Khan [16] presented a conceptual framework for an infant sleep detector based on face, blanket, movement, and eye detection, but their examination is performed only qualitatively, on infant dolls.

More modern deep learning methods have only recently emerged. In 2023, Choi et al. [7] use "(2+1)-dimensional" convolutional networks and sliding windows for action segmentation in closely-related task of sleep apnea detection, incorporating a massive dataset of 729 subjects of all ages. In the same year, Singh et al. [22] contributed another conceptual framework for infant sleep monitoring, this one based on infant pose estimation, but not tested on any video subjects. Finally, in 2024, Huang et al. [12] proposed a frame-based convolutional feature extractor with a fully connected classifier, and also made novel use of domain alignment for consistency across hospital environments. While this approach still does not use spatiotemporal deep learning, their study is notable for the sheer size of their dataset population, which includes 10–30 m videos of 103 term or preterm infants in the NICU. Their method also relies on a controlled setting with clean face detections, but they do achieve good sleep–wake classification results, including in cross-environment tests.

3. An End-to-End Sleep–Wake Classifier

Our proposed end-to-end sleep–wake classifier (see Figure 1) addresses the challenges of capturing nuanced spa-

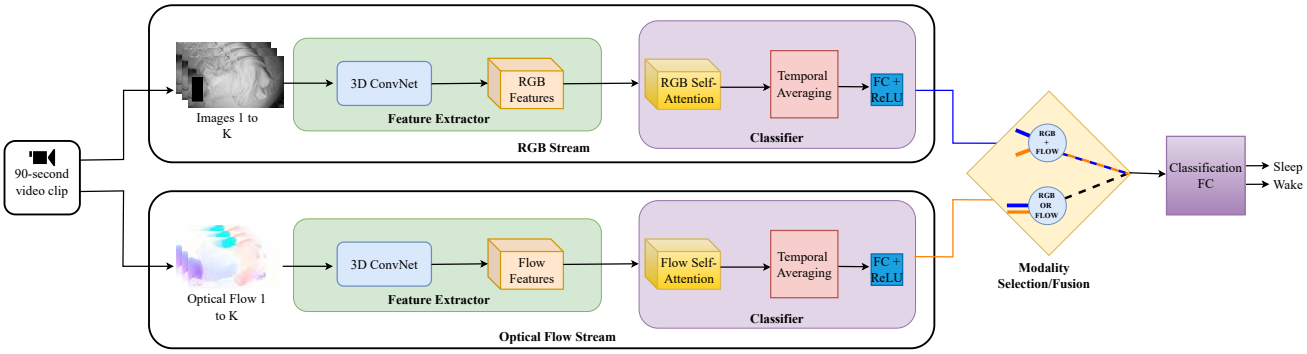

Figure 1. Architecture of our two-stream network for infant sleep–wake classification. The model processes 90-second video clips through parallel RGB and optical flow streams. Each stream consists of: (1) a Feature Extractor using 3D ConvNet to extract spatiotemporal features from K sequential frames, (2) a Classifier incorporating self-attention mechanisms for temporal dependency modeling, temporal averaging, and fully connected layers with rectified linear unit (ReLU) activation. A modality selection/fusion module allows flexible stream combination (RGB alone, flow alone, or RGB + flow) before final classification into Sleep/Wake states. The architecture highlights two main processing stages: feature extraction (green blocks) and classification (purple blocks).

tial and temporal patterns in infant sleep behavior by leveraging advanced feature extraction and classification techniques. Given the need to differentiate subtle and context-dependent sleep–wake states, we employ a two-stream architecture that combines complementary modalities: RGB frames for spatial information and optical flow for motion dynamics. The choice of Inflated 3D ConvNet (I3D) [6] as our backbone for feature extraction stems from its proven ability to capture spatiotemporal patterns in video data. The I3D model is pretrained on the Kinetics 400 [15] dataset. By integrating these features, our pipeline effectively captures both static postural cues and dynamic motion patterns, enabling a more comprehensive understanding of infant sleep–wake states.

3.1. Extracting RGB and Optical Flow Features

Given a 90 s video clip, we first sample K frames uniformly across the temporal dimension. Let $\mathbf{X}_{\text{RGB}} \in \mathbb{R}^{K\times H\times W\times 3}$ represent the RGB input tensor, where H and W denote the height and width of each frame. For optical flow computation, we use Recurrent All-Pairs Field Transforms for Optical Flow (RAFT) [25] to generate flow fields between consecutive frames, resulting in an input tensor $\mathbf{X}_{\text{flow}} \in \mathbb{R}^{K\times H\times W\times 2}$, where the last dimension represents horizontal and vertical motion components. The time separation between frames used for the optical flow calculation was 0.1 s, corresponding to the video frame rate of 10 frames per second. The I3D model (Figure 1, green box) processes these inputs through two parallel streams. Each stream consists of a series of 3D convolutions that maintain temporal information while learning spatial features. For a given input tensor $\mathbf{X}$, the feature extraction process can be expressed as:

$$\mathbf{F} = \text{I3D}(\mathbf{X}; \theta_{\text{I3D}}), \tag{1}$$

where θ_{I3D} represents the pre-trained model parameters, and $\mathbf{F} \in \mathbb{R}^{K'\times 1024}$ is the extracted feature matrix, with K' temporal steps and 1024 feature channels.

3.2. Self-Attention and Feature Fusion

The extracted features from each stream are independently processed through temporal self-attention modules to capture long-range dependencies. For each feature stream $\mathbf{F}$, we compute attention scores $\mathbf{A}$ via

$$\mathbf{Q} = \mathbf{F}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{F}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{F}\mathbf{W}_V, \tag{2}$$

and

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \tag{3}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are learnable parameter matrices, and d_k is the dimension of the key vectors. The attention outputs are temporally averaged to obtain a single feature vector per stream:

$$\mathbf{f}_{\text{RGB}} = \frac{1}{K'}\sum i = 1^{K'} \mathbf{A}^{(i)}_{\text{RGB}}; \quad \mathbf{f}_{\text{flow}} = \frac{1}{K'}\sum_{i=1}^{K'} \mathbf{A}^{(i)}_{\text{flow}}. \tag{4}$$

3.3. Sleep–Wake Classifier

Our two-stream fusion classifier (Figure 1, purple box) processes temporal features extracted from the I3D model. From each 90-second clip, we sample $K = 14$ frames uniformly distributed across time, chosen empirically to balance computational efficiency with temporal coverage. For each stream, the input features $\mathbf{F} \in \mathbb{R}^{K\times 1024}$ are processed through multi-head self-attention

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{F}\mathbf{W}_Q(\mathbf{F}\mathbf{W}_K)^T}{\sqrt{d_k}}\right)\mathbf{F}\mathbf{W}_V, \tag{5}$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, $\mathbf{W}_V$ are learnable parameters and d_k is the dimension of the key vectors. This self-attention mechanism enables the model to adaptively focus on salient temporal patterns while suppressing less relevant information.

The attention outputs are temporally averaged and transformed through independent fully connected layers with ReLU activation and dropout ($p = 0.25$):

$$\mathbf{h}_{\text{RGB,flow}} = \text{Dropout}\left(\text{ReLU}\left(\mathbf{W}\frac{1}{K}\sum_{i=1}^{K}\mathbf{A}^{(i)}_{\text{RGB,flow}} + \mathbf{b}\right)\right). \tag{6}$$

The two streams are fused via element-wise multiplication $\mathbf{h}_{\text{fused}} = \mathbf{h}_{RGB} \odot \mathbf{h}_{flow}$, and passed through a final classification layer to output sleep–wake probabilities:

$$\mathbf{p} = \text{softmax}(\mathbf{W}_{\text{cls}}\mathbf{h}_{\text{fused}} + \mathbf{b}_{cls}), \tag{7}$$

where $\mathbf{p} \in \mathbb{R}^2$ represents probabilities for sleep and wake states. The final prediction is determined by thresholding p_{wake} at 0.5. The model is trained end-to-end using binary cross-entropy loss and subject-wise data splits to ensure robust evaluation (see Section 4.4).

4. SmallSleeps: Our Infant Sleep Dataset

To develop a robust and scalable system for infant sleep–wake classification, we collected and annotated a comprehensive dataset of overnight video recordings captured in real-world home environments. Our dataset, SmallSleeps, comprises 152 hours of footage from 17 infants aged 4–11 months, recorded using video cameras set up by caregivers in their cribs. Behavioral coding, rather than physiological signals, was employed to annotate sleep and wake states, allowing for non-invasive data collection. This process addressed key challenges such as environmental variability, inconsistent lighting, and limited data diversity while ensuring a balanced representation of sleep, wake, and mixed-state transitions for model training and evaluation.

4.1. Data Collection

Overnight sleep video footage was collected and annotated by our clinical neurodevelopment team (see Figure 2 for the annotation tool setup), with Northeastern University Institutional Review Board approval (IRB #17-08-19). Participants were from 4–11 months old. Video cameras were sent to infants' homes and baby monitors were set up in cribs by caregivers and activated for overnight recordings, triggering the camera's monochromatic infrared mode.

A total of 152 hours of video footage was captured, from one night's sleep per participant. The videos were recorded at 10 frames per second. These data were collected as part of a study into infant sleeping and sucking behavior, and was previously used in our work on infant sucking detection [30, 31], but the behavioral coding and dataset preparation for the present sleep detection work is new.

Figure 2. Sleep–wake annotation using the VGG Image Annotator tool.

4.2. Annotations from Behavioral Coding

In order to capture a broad population of infant subjects, carrying out natural behavior in their own cribs at home, our study design forgoes the use of physiological sensors to monitor sleep. Our sleep state annotations are therefore derived from behavioral coding, performed by research assistants led and orchestrated by our neurobehavioral research team. Given the enormity of the task of (double) coding over 150 hours of footage, we developed a strategy in which each video was reviewed in near-real-time by two research assistants, who placed time markers for the start and end of waking-like states or behavioral units, including arousal, fussing, crying, and true waking. True wake states, which can be more complex and require minutes-long periods of observation and confirmation, were then inferred from the aggregation of both sets of codes. This removed the need for coders to continually backtrack and make subtle judgment calls.

More precisely, coders annotated videos for the aggregate behavioral units of arousal, defined as follows, and drawn from Symanski et al. [24], Holditch-Davis et al. [11], and, especially for the arousal behavior, Hayes [9], and Hayes et al. [10]:

Arousal is generally initiated by the opening of the eyes, often accompanied by head or body movement directed towards orientation to the environment. After initiation, the arousal event lasts until (and includes) a period of drowsiness, during which the eyes may open and close within a few seconds. Spontaneous movement during sleep, without the eyes opening, does not initiate arousal. Crying and fussing is common after the start of arousal and should be included in the arousal state. A start time stamp for arousal should be placed after 5 seconds of continuous arousal activity, and an end time stamp placed after 5 seconds of returning to sleep.

Non-Arousal is the default alternative state for the rest of the night, after initial onset of the night's sleep session, and before the final offset.

Afterwards, the two sets of arousal time segments from both coders were merged, and in addition, arousal segments separated by less than 180 s of non-arousal were merged into one long arousal segment, and isolated arousal events lasting longer than 60 s were deleted. The resulting, modified arousal-derived event segments were taking as the final ground truth segment labels for waking states, with the remaining time between the night's onset and offset labeled as sleep. These adjustments were performed after our neurobehavioral team analyzed the initial arousal coding. They were empirically judged to rectify coder discrepancies stemming from differences in interpretation of the arousal criteria, as well as omissions as a result of human error, and to yield inferred sleep–wake states in line with current scientific understanding. Video footage from two infants out of an original 19 were removed due to a lack of arousal activity, resulting in 17 final subjects for our study.

4.3. Sampling Short Video Clips

Our long-term goal is to develop a model capable of not only classifying, but also temporally segmenting sleep–wake states in long, naturalistic overnight recordings. Achieving this requires a sampling methodology that captures the realistic challenges of identifying mixed-state clips, which reflect transitions or overlaps between sleep and wake states. To address this and evaluate our model's performance with mixed-state clips, we designed a sampling strategy that samples both pure and mixed-state clips. We introduce five threshold bins to classify 90 s clips based on the percentage of time spent in the annotated state: 90–100%, 80–90%, 70–80%, 60–70%, 50–60%. For our final training set, we use clips in the 70–100% "purity" range. During overnight sleep, wake states were sparse and therefore act as the limiting factor for video clip sampling. For each subject, we repeatedly sample 90 s clips from the video until we reach the desired 100 clips per bin for the wake class. This process ensures that our sampled states include adequate representation of high-purity states (90–100%) as well as mixed-state clips where sleep activity is present alongside wake. The same process is repeated for sleep clips until 100 samples per bin are obtained.

4.4. Training and Test Sets

To evaluate the robustness of our model, we perform a subject-wise train–test split, with 10 subjects chosen at random for training and the remaining 7 used for testing. This split ensures that the test data comes from entirely unseen subjects, creating a realistic test of the model's ability to generalize to new environments and conditions. The use of unseen subjects in the test set allows us to assess whether the model is learning features that are generalizable across diverse conditions and avoid over-fitting to the specific characteristics of the training data.

5. Experimental Results

5.1. Implementation Details

Hyperparameters We use the feature extractor from I3D to generate the RGB and optical flow feature vectors, each with dimensions 1024×14, where 1024 represents the feature channels, and 14 corresponds to the temporal steps. These feature vectors are individually processed through a self-attention layer, and the outputs are aggregated using mean pooling, reducing each sequence to a single feature vector of dimensions 1024×1 per stream. Next, these vectors are passed through a hidden layer that reduces their dimensionality to 512×1. The resulting features are then passed through a fully connected layer, followed by ReLU activation and a dropout layer with a dropout probability of 0.25. The two-stream outputs are fused via element-wise multiplication to form a combined representation, which is passed to a classification layer to predict sleep or wake states. The model is optimized using the Adam optimizer with an initial learning rate of 10^{-5}. We integrate a `ReduceLROnPlateau` learning rate scheduler that halves the learning rate if the test loss does not improve for two consecutive epochs. Training is conducted for 50 epochs with a batch size of 16, using a subject-wise data split (see Section 4.4 for details). Our entire feature extraction and model training pipeline ran on Nvidia Tesla V100 GPUs.

Models We develop two models based on the purity composition of the datasets used for training. Note that, as per Section 4.4, both models draw from a uniform number of clips per bin and per state (sleep or wake).

Pure-State Model (90–100) Trained on 1,999 clips from the 90–100 threshold bin (1,000 sleep and 999 wake)

Mixed-State Model (70–100) Includes all clips from the pure-state dataset, plus 200 additional clips each from the 70–80 and 80–90 bins totaling 2,399 clips.

Validation is performed on a consistent dataset of 4,150 clips from 7 unseen subjects, with approximately 1,400 clips per threshold bin. This setup allows direct comparison between models trained solely on pure-state clips and those trained on a mix of pure and mixed-state clips.

5.2. Performance Comparison of Optical Flow and RGB Features

Table 2 reports the performance of various models under accuracy, precision, and recall metrics, stratified by the

Table 2. Comprehensive evaluation of our sleep–wake classification model with wake as the positive class, showing accuracy, precision, and recall on the overall test set, as well as broken down by test bin based on state purity threshold. The model was evaluated using different feature combinations (RGB, optical flow, and their fusion) and training datasets (clips from the 90–100 bin vs. the 70–100 bins combined). Best performing scores are highlighted in **bold**.

		Performance Overall			**Performance by Test Bin**								
		70–100			70–80			80–90			90–100		
Features	**Training Bin**	Acc	Prec	Rec	Acc	Prec	Rec	Acc	Prec	Rec	Acc	Prec	Rec
RGB	90–100	0.72	0.73	0.72	0.63	0.63	0.67	0.69	0.69	0.72	0.83	0.88	**0.86**
Optical Flow	90–100	0.76	0.73	**0.84**	0.69	0.65	**0.85**	0.72	0.68	**0.83**	**0.88**	0.90	0.85
RGB + Flow	90–100	0.77	0.74	**0.84**	0.70	0.66	**0.85**	0.73	0.70	0.82	**0.88**	0.90	0.84
RGB	70–100	0.71	0.72	0.70	0.63	0.63	0.67	0.68	0.69	0.68	0.83	0.88	0.76
Optical Flow	70–100	**0.78**	**0.76**	0.81	**0.72**	**0.69**	0.83	**0.75**	**0.73**	0.81	0.87	**0.91**	0.81
RGB + Flow	70–100	**0.78**	**0.76**	0.82	**0.72**	0.68	0.83	0.74	0.71	0.81	0.87	**0.91**	0.82

architectural and training configurations specified above. These results show that when using RGB features alone, the models achieve moderate accuracy, with the pure-state model yielding an accuracy of 71% and the mixed-state model achieving 72% (see Rows 1 and 4 in Table 2). Introducing optical flow features in addition to RGB, results in a substantial improvement in model performance, with the best model reaching an accuracy of 78% (see Row 6 in Table 2). Notably, for both pure- and mixed-state models trained with optical flow features, the recall increases by 12 and 11 points, respectively, while maintaining or slightly increasing precision. The most significant improvement is observed in the 70–80 bin, where recall is 18 points higher and precision is 3 points higher in the combined RGB + flow model compared to the RGB-only model. These results underscore the importance of optical flow in capturing subtle wake events, particularly in mixed-state clips, and it enables the model to more effectively differentiate between sleep and wake than RGB features alone.

The optical flow model exhibits significantly better sensitivity to wake events compared to the RGB-only model. This suggests that the extracted optical flow features, which capture temporal and motion-related cues, are more representative of wake behavior, particularly in ambiguous cases where the distinction between sleep and wake is less clear. Subtle movements, such as shifts in body position or eye movement, which might be overlooked in RGB frames, are effectively captured by optical flow features, demonstrating their relevance in distinguishing between sleep and wake states.

Given that the videos were recorded overnight in dimly lit conditions, generally activating the infrared mode yielding grayscale videos, the static information from RGB features is minimal and sometimes misleading. While useful for general visual recognition tasks, they lack the temporal context required for sleep classification. This is particularly problematic for wake events that often involve subtle motion, which may not be readily visible in individual RGB frames. As a result, the RGB-only model has difficulty detecting these small but significant wake signals, resulting in lower performance, particularly for mixed-state clips.

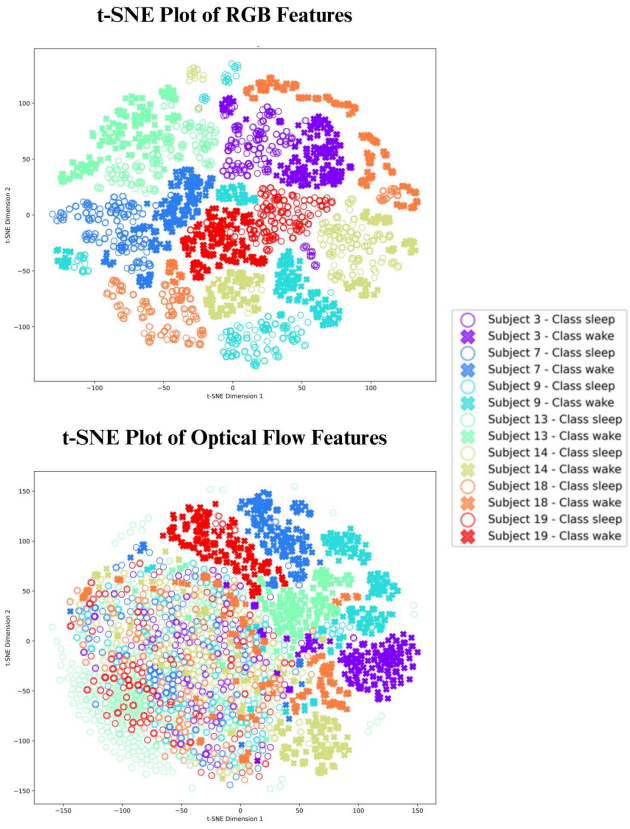

Figure 3. The t-SNE plots of RGB and optical flow features (taken from our feature extractor as seen in Figure 1) of video clips from the 90–100 bin from the test set. The seven different colors represent the seven test subjects, with a filled **X** marker representing the wake class and an open **O** marker representing the sleep class.

Further insights are provided by the t-distributed stochastic neighbor embedding (t-SNE) plots in Figure 3. In the optical flow t-SNE plot, distinct clusters for wake states are observed, predominantly positioned towards the left.

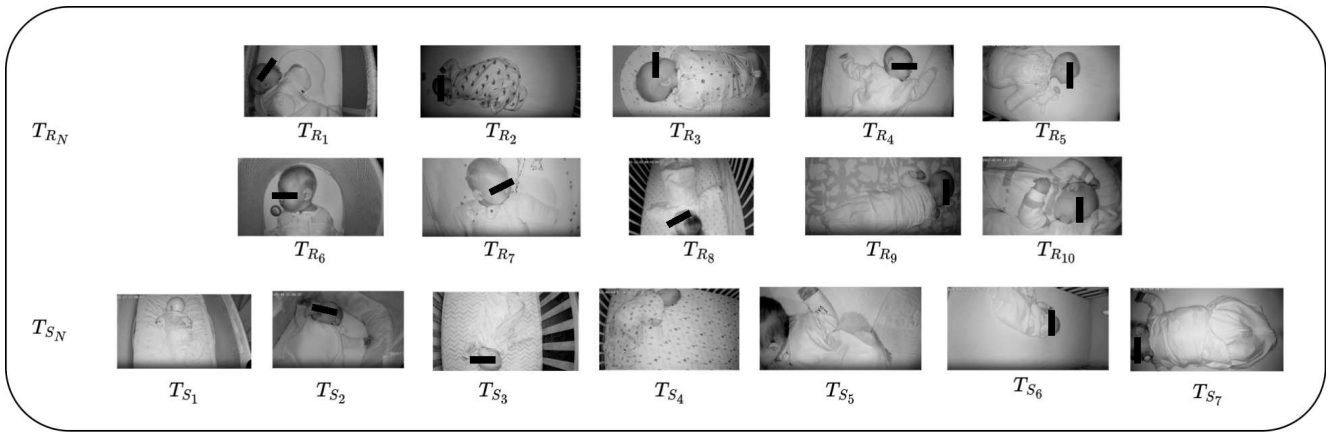

Figure 4. Distribution of subjects in our dataset. Top two rows (T_{R_N}): Training subjects ($N = 10$) showing representative frames from each infant. Bottom row (T_{S_N}): Test subjects ($N = 7$) demonstrating the independent test set. Subscript notation R denotes training subjects and S denotes test subjects, while the numerical index uniquely identifies each subject within their respective sets.

Each subject's features form separate, well-defined clusters, suggesting that optical flow captures subject-specific motion patterns effectively. This clustering indicates that motion patterns provide a unique signature for each subject during wake states. Conversely, sleep states form a broader, less distinct cluster towards the right of the plot. The lack of subject-specific clustering for sleep states may be due to the homogeneity of motion patterns across subjects during sleep, particularly in low-activity periods. In the RGB t-SNE plot, subject-specific clusters are visible, indicating that RGB features effectively capture the unique static spatial information, such as the visual appearance of individual subjects. However, within each subject's cluster, sleep and wake states overlap considerably. This overlap suggests that RGB features are less effective at distinguishing between these states, likely because they lack the temporal dynamics critical for differentiating motion-based wake behaviors.

While optical flow features excel in distinguishing wake states, RGB features provide complementary information that may not be fully represented in motion-based features. This synergy is reflected in the fused RGB + flow model's performance (Rows 3 and 6 in Table 2). The slight performance improvement may suggest that RGB features contribute additional context that enriches the model's ability to generalize across different states and subjects.

5.3. Subject-Wise Performance Comparison

A qualitative inspection of the dataset (see Figure 4) reveals interesting patterns in the relationship between video quality and classification performance. The faces of subjects where the algorithm performs best, such as T_{S_1}, T_{S_2}, and T_{S_3} appear prominently visible throughout the video. This may provide the model with more reliable motion cues, that describe the superior performance (97%, 95%, and 94% for T_{S_1}, T_{S_2}, and T_{S_3}, respectively). Conversely, for test subjects where algorithm performance is weaker, we notice the face becomes more obstructed and obscured, which could explain the inability to detect subtle movements in the facial region that may distinguish between states. Observe in Figure 4 where the face of subjects T_{S_5}, T_{S_6}, and T_{S_7} appear off-center, darkened, or out of frame. Additionally, the camera angle for lower-performing subjects (see T_{S_6} in Figure 4) appears more deviated, capturing the scene from a lower or skewed perspective rather than the ideal top-down view of the crib, representative of the better-performing subjects. These observations underscore the challenges inherent to small data research in infant sleep–wake classification. Variability in video conditions, such as differences in infant positioning, camera angles, and lighting, introduces additional complexity that strains the generalization capabilities of models trained on small, heterogeneous datasets. The qualitative differences observed across subjects exemplify how these constraints manifest in practice with the varying performance across our seven test subjects.

Table 3. Per-subject classification performance of our best performing model (RGB + flow trained with 70–100 data) across different threshold bins. Results show accuracy scores for each test subject (T_{S_1} through T_{S_7}) with the best-performing scores for each threshold highlighted in **bold**.

	Test Subject						
Test Bin	T_{S_1}	T_{S_2}	T_{S_3}	T_{S_4}	T_{S_5}	T_{S_6}	T_{S_7}
70–80	0.58	0.78	0.82	0.66	0.65	0.70	**0.89**
80–90	0.66	0.80	**0.89**	0.68	0.66	0.64	0.86
90–100	**0.97**	0.95	0.94	0.90	0.88	0.77	0.69

6. Conclusion and Future Work

In this study, we presented a novel spatiotemporal deep-learning framework for classifying infant sleep–wake states using video data. We introduced SmallSleeps, a dataset comprised of 152 hours of overnight footage of infants,

recorded using video cameras set up in their cribs. Our approach revealed the challenges of analyzing real-world overnight video data, including variability in environmental conditions and the subtleties of infant motion. Our results demonstrate the efficacy of optical flow features, particularly in enhancing performance on mixed-state clips, underscoring the importance of capturing motion-related cues for wake detection.

Future research on incorporating infant face and body pose estimation into the two-stream model to highlight specific eye movements or respiration patterns could lead to better performance, especially if failed detections or estimations can be handled sensibly. Ultimately, these enhancements aim to pave the way for a comprehensive and non-invasive system for infant sleep monitoring that could support both clinical and at-home applications.

References

[1] Saadullah Farooq Abbasi, Awais Abbas, Iftikhar Ahmad, Mohammed S Alshehri, Sultan Almakdi, Yazeed Yasin Ghadi, and Jawad Ahmad. Automatic neonatal sleep stage classification: A comparative study. *Heliyon*, 2023. 1

[2] HW Agnew Jr, Wilse B Webb, and Robert L Williams. The first night effect: An eeg studyof sleep. *Psychophysiology*, 2(3):263–266, 1966. 1

[3] Muhammad Awais, Xi Long, Bin Yin, Saadullah Farooq Abbasi, Saeed Akbarzadeh, Chunmei Lu, Xinhua Wang, Laishuan Wang, Jiong Zhang, Jeroen Dudink, and Wei Chen. A Hybrid DCNN-SVM Model for Classifying Neonatal Sleep and Wake States Based on Facial Expressions in Video. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1441–1449, May 2021. 3

[4] Natalie Barnett, Assaf Glazer, Tor Ivry, Yanai Ankri, and Haviva Veler. Computer vision algorithms outperform actigraphy. In *Paediatric Sleep Medicine*, page P134. European Respiratory Society, Apr. 2019. 2, 3

[5] S. Cabon, F. PorÃ©e, A. Simon, B. Met-Montot, P. Pladys, and N. Nardi. Audio- and video-based estimation of the sleep stages of newborns in Neonatal Intensive Care Unit. *Biomedical Signal Processing and Control*, 52:362–370, July 2019. 2, 3

[6] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. *CoRR*, abs/1705.07750, 2017. 4

[7] You Rim Choi, Gyeongseon Eo, Wonhyuck Youn, Hyojin Lee, Haemin Jang, Dongyoon Kim, Hyunwoo Shin, and Hyung-Sin Kim. SlAction: Non-intrusive, Lightweight Obstructive Sleep Apnea Detection using Infrared Video, Sept. 2023. arXiv:2309.02713 [cs]. 3

[8] Smadar Gertner, Charles W Greenbaum, Avi Sadeh, Zipora Dolfin, Leah Sirota, and Yocheved Ben-Nun. Sleep–wake patterns in preterm infants and 6 month's home environment: implications for early cognitive development. *Early human development*, 68(2):93–102, 2002. 1

[9] Marie J. Hayes. Methodological issues in the study of arousals and awakenings during sleep in the human infant. In *Awakening and sleep-wake cycle across development.*, Advances in Consciousness Research, vol. 38., pages 23–45. John Benjamins Publishing Company, Amsterdam, Netherlands, 2002. 5

[10] M J Hayes, M R Akilesh, M Fukumizu, A A Gilles, B A Sallinen, M Troese, and J A Paul. Apneic preterms and methylxanthines: arousal deficits, sleep fragmentation and suppressed spontaneous movements. *Journal of Perinatology*, 27(12):782–789, Dec. 2007. 5

[11] Diane Holditch-Davis, Mark Scher, Todd Schwartz, and Diane HudsonâBarr. Sleeping and waking state development in preterm infants. *Early Human Development*, 80(1):43–64, Oct. 2004. 5

[12] Dongmin Huang, Dongfang Yu, Yongshen Zeng, Xiaoyan Song, Liping Pan, Junli He, Lirong Ren, Jie Yang, Hongzhou Lu, and Wenjin Wang. Generalized Camera-Based Infant Sleep-Wake Monitoring in NICUs: A Multi-Center Clinical Trial. *IEEE Journal of Biomedical and Health Informatics*, pages 1–14, 2024. 3

[13] Xiaofei Huang, Lingfei Luan, Elaheh Hatamimajoumerd, Michael Wan, Pooria Daneshvar Kakhaki, Rita Obeid, and Sarah Ostadabbas. Posture-based Infant Action Recognition in the Wild with Very Limited Data. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4912–4921, Vancouver, BC, Canada, June 2023. IEEE. 2

[14] Conrad Iber. The aasm manual for the scoring of sleep and associated events: rules, terminology, and technical specification. *(No Title)*, 2007. 1

[15] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 4

[16] Tareq Khan. An Intelligent Baby Monitor with Automatic Sleeping Posture Detection and Notification. *AI*, 2(2):290–306, June 2021. 3

[17] Xi Long, RenÃ©e Otte, Eric Sanden, Jan Werth, and Tao Tan. Video-Based Actigraphy for Monitoring Wake and Sleep in Healthy Infants: A Laboratory Study. *Sensors*, 19(5):1075, Mar. 2019. 3

[18] Yohei Mukai, Kento Morita, Nobu C. Shirai, Tetsushi Wakabayashi, Harumi Shinkoda, Asami Matsumoto, Yukari Noguchi, and Masako Shiramizu. Automatic Classification of Neonatal Sleep-Wake States Based on Facial Video Analysis. In *2019 4th International Conference on Information Technology Research (ICITR)*, pages 1–6, Moratuwa, Sri Lanka, Dec. 2019. IEEE. 3

[19] Howard P Roffwarg, Joseph N Muzio, and William C Dement. Ontogenetic development of the human sleep-dream cycle: The prime role of" dreaming sleep" in early life may be in the development of the central nervous system. *Science*, 152(3722):604–619, 1966. 1

[20] Avi Sadeh, M. Sharkey, and Mary A. Carskadon. Activity-Based Sleep-Wake Identification: An Empirical Test of Methodological Issues. *Sleep*, 17(3):201–207, May 1994. 2, 3

[21] A. J. Schwichtenberg, Jeehyun Choe, Ashleigh Kellerman, Emily A. Abel, and Edward J. Delp. Pediatric Videosomnography: Can Signal/Video Processing Distinguish Sleep and Wake States? *Frontiers in Pediatrics*, 6:158, June 2018. 2, 3

[22] Gurpreet Singh, Abhishek Raj Shekhar, Xinrui Yu, and Jafar Saniie. Smart Infant Monitoring System Using Computer Vision and AI. In *2023 IEEE International Conference on Electro Information Technology (eIT)*, pages 1–6, Romeoville, IL, USA, May 2023. IEEE. 3

[23] Ambra Stefani, David Gabelia, Thomas Mitterling, Werner Poewe, Birgit Högl, and Birgit Frauscher. A prospective video-polysomnographic analysis of movements during physiological sleep in 100 healthy sleepers. *Sleep*, 38(9):1479–1487, 2015. 1

[24] Mary Ellen Symanski, Marie J Hayes, and M Kumar Akilesh. Patterns of premature newborns' sleep-wake states before and after nursing interventions on the night shift. *Journal of Obstetric, Gynecologic, & Neonatal Nursing*, 31(3):305–313, 2002. 5

[25] Zachary Teed and Jia Deng. RAFT: recurrent all-pairs field transforms for optical flow. *CoRR*, abs/2003.12039, 2020. 4

[26] Olivia Walch, Yitong Huang, Daniel Forger, and Cathy Goldstein. Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. *Sleep*, 42(12):zsz180, 2019. 1

[27] Michael Wan, Shaotong Zhu, Lingfei Luan, Prateek Gulati, Xiaofei Huang, Rebecca Schwartz-Mette, Marie Hayes, Emily Zimmerman, and Sarah Ostadabbas. InfAnFace: Bridging the Infant–Adult Domain Gap in Facial Landmark Estimation in the Wild. *26th International Conference on Pattern Recognition (ICPR)*, 2022. 2

[28] Bernice M Wulterkens, Pedro Fonseca, Lieke WA Hermans, Marco Ross, Andreas Cerny, Peter Anderer, Xi Long, Johannes P van Dijk, Nele Vandenbussche, Sigrid Pillen, et al. It is all in the wrist: Wearable sleep staging in a clinical population versus reference polysomnography. *Nature and Science of Sleep*, pages 885–897, 2021. 1

[29] Yongshen Zeng, Xiaoyan Song, Liping Pan, Junli He, Lirong Ren, Jie Yang, Hongzhou Lu, and Wenjin Wang. Generalized camera-based infant sleep-wake monitoring in nicus: A multi-center clinical trial. *IEEE Journal of Biomedical and Health Informatics*, page 1–14, 2024. 1

[30] Shaotong Zhu, Michael Wan, Elaheh Hatamimajoumerd, Cholpady Vikram Kamath, Kashish Jain, Samuel Zlota, Emma Grace, Cassandra Rowan, Matthew Goodwin, Rebecca Schwartz-Mette, Emily Zimmerman, Marie Hayes, and Sarah Ostadabbas. A video-based end-to-end pipeline for non-nutritive sucking action recognition and segmentation in young infants. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 10 2023. 5

[31] Shaotong Zhu, Michael Wan, Sai Kumar Reddy Manne, Elaheh Hatamimajoumerd, Marie J. Hayes, Emily Zimmerman, and Sarah Ostadabbas. Subtle signals: Video-based detection of infant non-nutritive sucking as a neurodevelopmental cue. *Computer Vision and Image Understanding*, 247:104081, Oct. 2024. 2, 5