ALPEC: A Comprehensive Evaluation Framework and Dataset for Machine Learning-Based Arousal Detection in Clinical Practice

Stefan Kraft STEFAN.KRAFT@IT-DESIGNERS.DE
IT-Designers Gruppe & University of Tübingen, Germany

Andreas Theissler HTTPS://ORCID.ORG/0000-0003-0746-0424
Aalen University of Applied Sciences, Germany

Vera Wienhausen-Wilke V.WIENHAUSEN-WILKE@KLINIKUM-ESSLINGEN.DE
Klinikum Esslingen, Germany

Philipp Walter PHILIPP.WALTER@IT-DESIGNERS.DE
IT-Designers Gruppe, Germany

Gjergji Kasneci GJERGJI.KASNECI@TUM.DE
Technical University of Munich, Germany

Hendrik Lensch HENDRIK.LENSCH@UNI-TUEBINGEN.DE
University of Tübingen, Germany

Abstract

Detecting arousals during sleep is crucial for diagnosing sleep disorders, yet the adoption of Machine Learning (ML) in clinical practice is hindered by a mismatch between clinical protocols and ML methods. Clinicians typically annotate only arousal onsets, whereas ML approaches conventionally rely on annotations for both the beginning and end. Moreover, no standardized evaluation methodology exists that is tailored to the specific needs of arousal detection in clinical practice. We address these challenges by proposing a novel post-processing and evaluation framework – Approximate Localization and Precise Event Count (ALPEC) – which optimizes arousal detectors to reflect operational priorities. We further advocate focusing on arousal onset detection and assess the impact of this on current training and evaluation schemes, addressing associated simplifications and challenges. Finally, we introduce the novel Comprehensive Polysomnographic (CPS) dataset that reflects the aforementioned clinical annotation constraints and includes modalities absent from existing datasets, demonstrating the benefits of leveraging multimodal data for arousal onset detection. Our contributions significantly advance the integration of ML-based arousal detection into clinical settings, narrowing the gap between technological advancements and clinical requirements.

Data and Code Availability This paper introduces the CPS dataset (Kraft et al., 2024; Goldberger et al., 2000) which we collected during clinical practice from 2021-2022 monocentrically at Klinikum Esslingen, Germany. It is released on the PhysioNet platform and is accessible under the PhysioNet Credentialed Health Data License 1.5.0. We also utilize the 2018 PhysioNet Challenge Dataset (Ghassemi et al., 2018; Goldberger et al., 2000) which is also available on the PhysioNet repository.

While the code for our training and evaluation procedures is proprietary, besides formalizing ALPEC (Section 3.2), we provide a schematic scheme comparison (Appendix E), pseudo-code (Appendix F), and instructions for running baseline training and evaluation schemes (Sections 3.1 and 3.3). The code for these baselines is publicly available[1]. The official documentation of the CPS dataset on PhysioNet as well as the supplementary material include Croissant (Akhtar et al., 2024) specifications, code, and instructions for data loading.

Institutional Review Board (IRB) Our study protocol was approved by the ethics committee of the Landesärztekammer Baden-Württemberg, Germany, on 2020-10-21 (committee number F-2020-105).

1. DeepSleep 2.0: https://github.com/rfonod/deepsleep2 (MIT license); sktime: https://doi.org/10.5281/zenodo.3749000 (BSD-3-Clause license)

1. Introduction

Arousals are short-term biological activation processes during sleep and wakefulness that elevate the organism from a lower to a higher state of mental and physical activity (Raschke and Fischer, 1997). Frequent arousals during sleep disrupt deep sleep stages and REM sleep, compromising the restorative function of sleep and causing fragmentation (Wetter et al., 2012). Arousals are indicative of several sleep disorders, with obstructive sleep apnea (OSA) being the most prevalent breathing-related sleep disorder. OSA, characterized by partial or complete obstructions of the upper airway, results in oxygen desaturation and frequent arousals (Wetter et al., 2012). This disorder is a significant public health concern, with prevalence estimates around 20% in men and 17% in women, and is linked to severe health outcomes like hypertension, cardiovascular disease, and increased mortality risk (Franklin and Lindberg, 2015; Wetter et al., 2012; Punjabi, 2008).

Detecting arousals is a routine task in polysomnographic (PSG) examinations, which involve comprehensive recording and analysis of various physiological parameters such as brain waves, blood oxygen levels, breathing, eye and leg movements during sleep, conducted in sleep laboratories. However, the diversity of equipment, software, and protocols across laboratories poses significant challenges for developing universally applicable Machine Learning (ML) models for arousal detection (Anido-Alonso and Alvarez-Estevez, 2023). Even among laboratories using the same equipment, differences in settings and protocols complicate the development of general-purpose ML-based detectors. The lack of large-scale time series datasets, absence of clear evaluation metrics, and limited consensus on theoretical and practical understanding of time series further impede progress (Garza and Mergenthaler-Canseco, 2023). A significant drop in performance of current sleep stage classification models on data from different sleep laboratories highlights the need for arousal detection models trained on data specific to the clinical environments where they will be used (Anido-Alonso and Alvarez-Estevez, 2023).

According to the American Association of Sleep Medicine (AASM), arousals are defined as abrupt changes in EEG frequency lasting at least 3 seconds following 10 seconds of sleep (Berry et al., 2012). Although research suggests that the duration of arousals may have clinical significance (Schwartz and Moxley, 2006; Shahrbabaki et al., 2021), this finding has not been integrated into clinical practice. This is evident from the AASM guidelines, which recommend reporting solely the number of arousals and the arousal index, a measure that quantifies the frequency of arousals per hour of sleep, for PSG diagnostics (Berry et al., 2012). This aligns with our CPS dataset, where nearly all annotated arousals are three seconds in duration, which is the default setting of the clinical annotation software. This limitation implies that only the onset annotations are practically useful. Current training approaches for arousal detection, which rely on full annotations encompassing both the start and end of each event, diverge from this clinical practice. Consequently, they are misaligned with the clinical reality, hindering their application in real-world settings.

Our **first main contribution** is advocating for a shift in focus from full event detection to detecting arousal onsets to better align with clinical needs. We explore the implications on various training methodologies for binary event detection, addressing both simplifications and emerging challenges. Conversely, aligning clinical practices with ML model requirements without evident patient benefits would unnecessarily burden sleep laboratories, increasing operational challenges such as the scarcity of trained scorers and long patient wait times.

In addition to task misalignment, the fragmented landscape of evaluation methodologies lacks the operational utility required for real-world healthcare settings. Our **second main contribution** is aligning the performance evaluation of arousal detection systems for decision support rather than autonomous decision-making for which we advocate using the F2 score. A clinical decision support system (CDSS) augments the clinical workflow by highlighting potential arousal occurrences, enabling clinicians to focus on the most relevant segments of sleep recordings. Clinicians still retain final responsibility for diagnostic conclusions, but the system streamlines their work by pinpointing events needing review. This approach adheres to ethical standards in healthcare (Fawzy et al., 2023) and anticipates evolving regulatory requirements, such as the forthcoming EU AI Act, which mandates human oversight for AI systems in critical areas like healthcare (Madiega, 2021).

This aligned evaluation approach is integral to ALPEC – a post-processing and performance evaluation framework that constitutes our **third main contribution**. ALPEC is embedded within a rel-

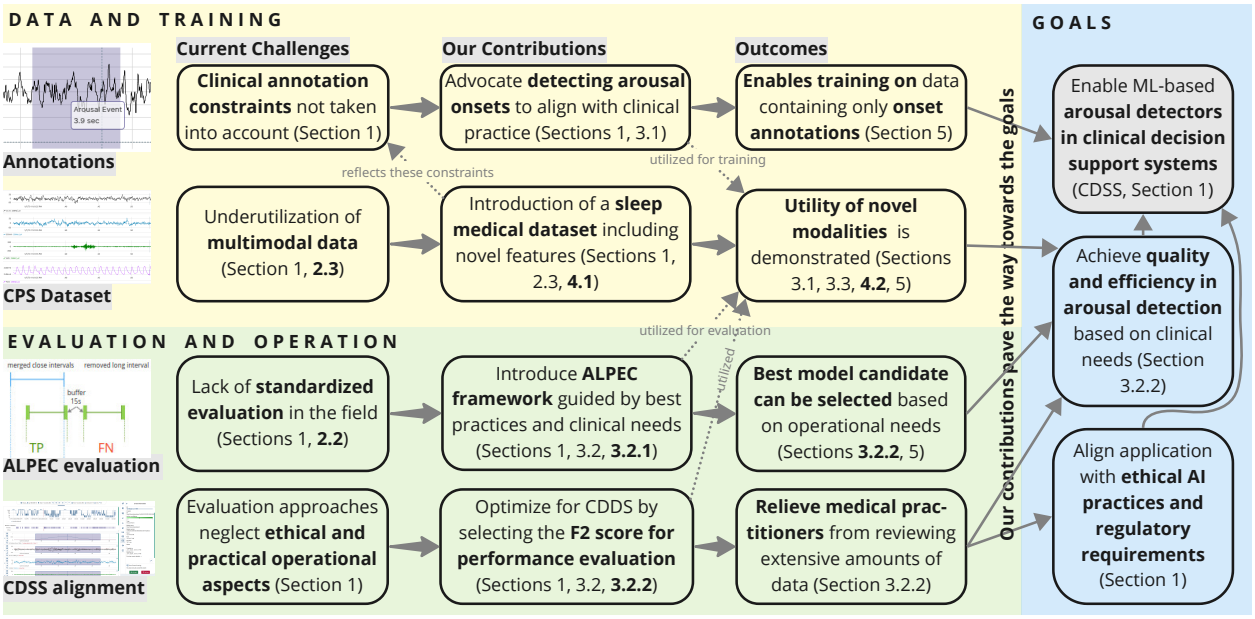

Figure 1: **Contextualization of our contributions** within the broader landscape of challenges in the field, highlighting the outcomes and underlying goals of our work. Key sections for each topic are referenced.

evant taxonomy, overcomes key conceptual limitations, and adheres to best practices. It is the first framework designed to standardize performance evaluation in the field of arousal detection, and it is tunable to the needs of other applications and domains that require precise event count detection. We provide a thorough comparison of our framework with existing evaluation methodologies.

Lastly, as our **fourth main contribution**, we are excited to introduce the Comprehensive Polysomnography (CPS) dataset, a curated, feature-rich collection unique for its extensive channels and novel beat-by-beat blood pressure annotations. This dataset aims to advance ML models in sleep disorder diagnostics. We present the first study on this dataset, demonstrating enhanced arousal detection capabilities through multimodal data, advancing the integration of previously underutilized data modalities.

Figure 1 provides a structured contextualization of our contributions and may serve as a guide for navigating the paper.

2. Related work

In this section, we embed our work into related work on arousal detection, highlighting training approaches (Section 2.1), diverse evaluation practices (Section 2.2), and notable datasets (Section 2.3).

2.1. Methods for arousal detection

Table 1 provides an overview of various approaches for arousal detection, highlighting its close association with sleep stage classification, where the primary goal is to determine the sleep stage for each 30-second epoch of a polysomnographic recording. Similar to sleep stage classification, for arousal detection, the data is typically segmented into N consecutive windows of fixed length s, with s either optimized as a hyperparameter or fixed at 30 seconds (Li et al., 2018; Phan et al., 2019) or other durations (Kuo et al., 2023). Overlapping windows are frequently employed to better capture arousal events, sometimes extended for evaluation as well (Badiei et al., 2023; Li et al., 2018). The definition of when a window signifies an arousal event often introduces another layer of complexity, sometimes determined by majority voting within the window or by the presence of at least one arousal label (Kuo et al., 2023), which may lead to another hyperparameter (Li et al., 2018).

We perform window-based segmentation only for baseline comparisons, as our primary focus is on continuous segmentation. For this, we build on the

Table 1: **Comparison of Related Work**. This comparison showcases the diversity of methodologies utilized in the field. The datasets are explained in Section 2.3. For counting *#Modalities*, multiple EEG channels are considered a single modality, while derived channels or features are counted separately. Notations: *Seg.* denotes segmentation methods; *(AU)PRC* and *(AU)ROC* indicate that either the curve, the area under the curve, or both are reported.

Authors	Task		Seg.		Dataset					Evaluation measures									Modalities					
	Arousal	Sleep Stage	Windowed	Pointwise	2018 Phys	SHHS	MESA	CPS (our)	Unpublished	(AU)PRC	(AU)ROC	Accuracy	Sensitivity	Specificity	Precision	Recall	β for F_β	Cohen's κ	EEG	EMG	ECG	EOG	Other	#Modalities
Badiei et al. (2023)	**x**	**x**	x		x	x					x	x	x						x		x			2
Foroughi et al. (2023)	x		x		x						x	x	x	x					x					1
Li et al. (2018)	x		x		x					**x**	**x**								x	x				2
Kuo et al. (2023)	x		x						x	x	x	x			x	x	1						x	18
Miller et al. (2018)	x			**x**	x					**x**	**x**								x	x	x		x	7
Howe-Patterson et al. (2018)	**x**	**x**		**x**	x					**x**	**x**								x	x		x	x	7
Li and Guan (2021)	**x**	**x**		**x**	x	x				**x**	**x**								x	x	x	x	x	8
Fonod (2022)	x			**x**	x					**x**	**x**								x	x	x	x	x	8
Zan and Yildiz (2023)	**x**	**x**		**x**		x	x			x	x	x			x	x	1	x	x					1
Ehrlich et al. (2024)	x			**x**		x			x	x							1		x	x		x		3
our	x		x	**x**	x			**x**							x	x	**2**		x	x	x	x	x	**39**

methodological foundation of the *DeepSleep* architecture, which employs a Fully Convolutional Neural Network (FCN) with a U-Net architecture to process extensive polysomnographic signals continuously (Li and Guan, 2021). This model differs from windowing approaches as it handles the entire dataset as a single sequence, where each point is evaluated within the context of its receptive field. This comprehensive approach to arousal detection offers several advantages over traditional window-based methods: It eliminates the need for multiple hyperparameters, does not require manual feature extraction, supports an end-to-end process, processes multimodal data natively, and leverages extensive temporal contexts to capture interactions across various timescales (Li and Guan, 2021). This has spurred a growing body of research pursuing similar comprehensive methodologies, as documented in Table 1.

2.2. Current state of evaluating arousal detection models

Evaluating arousal detection models is challenging due to the diversity of methodologies (e.g., pointwise vs. window-based evaluations) and the absence of standardized evaluation protocols (Foroughi et al., 2023; Badiei et al., 2023). This diversity – reflected in Table 1 – is compounded by the wide range of performance metrics employed. In practice, window-based evaluations dominate, as training typically favors window-based classification (WBC) over continuous segmentation (CS). For example, Zan and Yildiz (2023) use CS for both sleep stage classification and arousal detection in a multitask setup, yet still evaluate by applying 30-second non-overlapping sliding windows with labels assigned via majority vote or the mere presence of an arousal indicator. We employ pointwise and window-based evaluation methods only as baselines. In contrast, we propose an event-based evaluation approach that treats each contiguous segment of predictions as a single event to be compared with the ground-truth. Our method improves the segment-wise f-score introduced by Hundman et al. (2018) for time series anomaly detection (TSAD) and is also applicable when only event onsets are annotated. Notably, Ehrlich et al. (2024) also employ event-based evaluation for arousal detection with similar strategies – such as label adaptations, event merging, temporal tolerance, and a

specific counting scheme. However, our contribution goes further: Our approach (ALPEC) is embedded within a relevant taxonomy of evaluation metrics that was proposed by Sørbø and Ruocco (2023) for TSAD, which advocates tailoring the selection of evaluation metrics to operational needs. ALPEC addresses the requirements of clinical decision support systems and may serve to standardise the evaluation of arousal detectors, ending methodological fragmentation and ensuring comparability of results across studies.

2.3. Datasets and data modalities for arousal detection

Prominent datasets in sleep research include the 2018 PhysioNet Challenge dataset (Ghassemi et al., 2018; Goldberger et al., 2000), the Sleep Heart Health Study (SHHS) (Quan et al., 1997; Zhang et al., 2018), and the Multi-Ethnic Study of Atherosclerosis (MESA) (Chen et al., 2015; Zhang et al., 2018). Another notable and more recent collection of pediatric sleep data is the NCH Sleep DataBank (Lee et al., 2021, 2022; Goldberger et al., 2000). Like our CPS dataset, these datasets offer extensive PSG data and (in case of SHHS and MESA) patient information collected through standardized questionnaires such as sleep and restless legs questionnaires, the Pittsburgh Sleep Quality Index (PSQI), and the Epworth Sleepiness Scale (ESS). As indicated in the last column of Table 1, many studies utilize only few modalities. Those that employ more, such as Kuo et al. (2023) on an unpublished dataset, often perform feature engineering to derive additional features. Others, like Li and Guan (2021) and Fonod (2022), are constrained by the number of available channels in the 2018 PhysioNet Challenge dataset.

Our CPS dataset, in contrast, offers 17 raw channel modalities and numerous derived features by use of the DOMINO expert software from SOMNOmedics GmbH, featuring innovative modalities such as pulse transit time and beat-by-beat blood pressure estimations. The potential of these modalities in sleep diagnostics is supported by various studies (Misaka et al., 2020; Argod et al., 1998; Pitson et al., 1998, 1994), and we aim to further investigate their impact on arousal detection in an ongoing clinical study (Wienhausen-Wilke and Kraft, 2024). Additional highlights of the CPS dataset include annotations indicating whether arousals were first detected in the EEG or as a consequence of other physiological changes, along with detailed medical outcomes such as sleep diagnoses, Baveno classification, and T90 value. Further details on the CPS dataset are provided in Appendix H.

3. Methodologies

We present our core methodologies for event detection and performance evaluation, grounded in real-world considerations to enable robust clinical decision support for arousal detection in sleep medicine. We first describe our main approach for arousal onset detection in Section 3.1, present our novel framework ALPEC in Section 3.2, and then introduce multiple baseline approaches in Section 3.3.

3.1. Arousal detection by continuous segmentation

We start by detailing our main approach for arousal onset detection. We adopt the DeepSleep architecture, which facilitates continuous segmentation of data into distinct classes (Li and Guan, 2021). We build on an optimized version of this architecture, proposed by Fonod (2022), under MIT license, which reduces the U-Net's depth from 11 to 5 layers, substantially decreasing computational demands while maintaining comparable performance. This streamlined model processes all data points from multi-channel sleep recordings simultaneously, eliminating the need for window-based classification. It translates these inputs into sleep arousal scores for each data point using a binary cross-entropy (BCE) loss function. We refine this approach by employing a weighted BCE loss, adjusting the loss contribution of each data point by inversely weighting it according to the frequency of the arousal class within the subject's data, addressing class imbalance.

Since detecting singular arousal onset points does not work well with the DeepSleep approach (see Section 4.1), we modify the ground-truth annotations to mark intervals of length l around each arousal onset as positive. We select $l = 10$ seconds, aligning with arousal scoring rules that require at least 10 seconds of stable sleep between distinct arousal events, ensuring the created ground-truth intervals do not overlap (Berry et al., 2012). We call this approach *interval-based onset detection*. During inference, the DeepSleep model outputs probability scores $p_i(\mathbf{x})$ for each data point i. To smooth these outputs for reducing false detections, we apply an averaging filter over a smoothing window of $w = 3$ seconds per point.

3.2. ALPEC: Approximate localization and precise event count framework for post-processing and performance evaluation

Sørbø and Ruocco (2023) rightly state that there is no universally correct set of metrics for any specific task; however, using inappropriate metrics can lead to suboptimal decisions when selecting algorithms for productional use.

Guided by their taxonomy, we developed the Approximate Localization and Precise Event Count (ALPEC) framework to address the need for standardized performance evaluation in arousal detection that align with the operational goals of clinical practice. This is crucial for making informed decisions about the deployment of arousal detection algorithms in real-world clinical settings.

We first formally introduce the ALPEC procedure in Section 3.2.1, before discussing its rationale and our hyperparameter choice in Section 3.2.2. A schematic embedding of ALPEC into various training and evaluation schemes is provided in Appendix E (Figure 2), an algorithmic description in Appendix F, and a table of notation in Appendix A (Table 9).

3.2.1. Formal description of ALPEC

We now formally introduce the procedure of our post-processing and performance evaluation framework ALPEC.

ALPEC is compatible with both window-based classification (WBC) and continuous segmentation (CS) approaches to arousal detection. When using CS (see Section 3.1), we start with a probability score $p_{i,\nu}(\mathbf{x})$ for each data point $i = 1, ..., n$ from the measurement data $\mathbf{x}$ and each subject ν with $\nu = 1, ..., |D|$ in the dataset D, where $n = 2^{23}$ is the padded fixed number of data points for each input channel. Alternatively, if we use WBC (see Section 3.3), we start with probability scores $p_{\eta,\nu}(\mathbf{x})$ or binary class predictions $c_{\eta,\nu}(\mathbf{x})$, where $\eta = 1, ..., N$, and the data is divided into N windows of equal length s.

When we have probability scores, i.e., $p_{\xi,\nu}(\mathbf{x})$ with $\xi \in \{i, \eta\}$, we apply a threshold t_k to the scores to obtain binary class predictions $c_{\xi,\nu,k}(\mathbf{x})$, where thresholds t_k are selected from 0 to 1 in steps of 0.01, i.e., $k = 0, 1, ..., 100$:

$$c_{\xi,\nu,k}(\mathbf{x}) = \begin{cases} 1 & \text{if } p_{i,\nu}(\mathbf{x}) \geq t_k \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

In the case of WBC, we resample the window-based predictions to pointwise predictions $c_{i,\nu}(\mathbf{x})$ by assigning the prediction of the window to all data points within the window:

$$c_{i,\nu}(\mathbf{x}) = c_{\eta,\nu}(\mathbf{x}) \text{ for } i \in [(\eta - 1) \cdot s + 1, \eta \cdot s] \tag{2}$$

At this point, both starting points (continuous segmentation and window-based classification) are synchronous.

Interval Merging Next, we merge predictions less than δ seconds apart. For ease of notation, we temporarily drop the indices k and ν and parameter $\mathbf{x}$. At first, for the sequence of binary target values $C = (c_1, c_2, ..., c_n)$, we identify the start and end indices of each predicted interval P in C as P^{start} and P^{end}, respectively, where an interval starts at index i if $c_i = 1$ and $c_{i-1} = 0$ and ends at index j if $c_j = 1$ and $c_{j+1} = 0$ if all intermediate predictions $c_{i+1}, ..., c_{j-1}$ are equal to 1. For *arousal onset detection*, the distance is calculated based on the maxima of the scores of two consecutive predicted intervals. For each identified interval P, we find the index m within the interval that maximizes the score p_m:

$$m = \arg\max_{m \in [P^{\text{start}}, P^{\text{end}}]} s_m \tag{3}$$

Two intervals P_1 and P_2, with maximum score indices m_1 and m_2, are merged if $|m_1 - m_2| < \delta \cdot f$, where f is the sampling frequency of the data. In the case of *full event detection*, we merge two intervals based on their start and end points, i.e., if $|P_1^{\text{start}} - P_2^{\text{end}}| < \delta \cdot f$. The merged interval P^{merged} then extends from P_1^{start} to P_2^{end}. For the merged sequence C^{merged} we set:

$$c_i^{\text{merged}} = \begin{cases} 1 & \text{if } i \in \text{any} P^{\text{merged}} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Matching Predictions and Ground-Truth The final step is to compare the predicted intervals P with the ground-truth intervals G with start end points G^{start} and G^{end}. In the case of *point-based onset detection*, $G^{\text{start}} = G^{\text{end}}$, meaning the ground-truth intervals are points in time. Our ALPEC framework, however, remains generic and can be used for any length of ground-truth intervals, thus supporting both *interval-based onset detection* and *full event detection*.

Next, we once again tailor the evaluation method to the specificities of the task by introducing two two

key **approximate localization** components: First, we define a **maximum duration** d for the predicted intervals, which we will utilize shortly. Second, we extend all ground-truth intervals with a **temporal tolerance buffer** b^{before} on the left and b^{after} on the right side of the interval:

$$G^{\text{start,ext}} = \max(0, G^{\text{start}} - b^{\text{before}} \cdot f) \tag{5}$$

$$G^{\text{end,ext}} = \min(n, G^{\text{end}} + b^{\text{after}} \cdot f) \tag{6}$$

Now, we can compare predicted and ground-truth intervals to calculate true positive (TP), false positive (FP), and false negative (FN) counts, meeting the requirement of **precise event counts**. A TP is counted when a predicted interval P overlaps with a ground-truth interval G^{ext} extended by the buffer, so that $P^{\text{start}} \leq G^{\text{end,ext}}$ and $P^{\text{end}} \geq G^{\text{start,ext}}$ and $P^{\text{end}} - P^{\text{start}} \leq d \cdot f$, restricting the maximum duration of a predicted interval. A FP is counted if a P does not overlap with any G^{ext}. A FN is counted if a G^{ext} does not overlap with any P. If multiple P overlap with a single G^{ext}, we count one TP for the first overlap and each additional overlap as a FP. If a single P overlaps with multiple G^{ext}, we count one TP for the first match and each unmatched G^{ext} as a FN.

Performance Evaluation Next, we use established formulas to calculate the $\text{Precision}_{\nu,k}$, $\text{Recall}_{\nu,k}$, and $\text{F2}_{\nu,k}$ score for each subject ν and each threshold t_k from the TP, FP, and FN counts, reintroducing the indices.

We use the F2 score for selecting the optimal threshold. The first step is to calculate the micro average $F2_k^{\text{train}}$ score across subjects $v^{\text{train}} = 1, ..., |T|$ of the training set T for each threshold t_k. as $\text{F2}_k^{\text{train}} = \frac{1}{|T|} \sum_{\nu \in v^{\text{train}}} \text{F2}_{\nu,k}$. From this, we determine the optimal threshold t_k^{opt} with $k^{\text{opt}} = \arg\max_k \text{F2}_k^{\text{train}}$ which maximizes the average $F2$ score on the training set. We then obtain the mean Precision, Recall, and F2 score across subjects for the optimal threshold t_k^{opt}. We select the mean F2 score as the final metric for performance evaluation and report the mean Precision and Recall as auxiliary metrics for additional insights (performance analysis).

3.2.2. Rationale and Hyperparameter Choice for Arousal Detection

ALPEC shares several similarities with existing metrics. First, it is most similar to the segment-wise f-score (Hundman et al., 2018), as both approaches focus on evaluating segment overlaps rather than pointwise predictions. Second, like some existing methods, ALPEC employs a **temporal tolerance buffer** around ground-truth intervals (Scharwächter and Müller, 2020), that we set to $b^{\text{before}} = b^{\text{after}} = 15s$, corresponding to the typical length of one 30-second epoch as viewed by medical scorers. The use of this buffer addresses potential temporal inaccuracies in arousal event annotations and is backed by the irrelevance of precise annotations in current clinical practice, leading to the **approximate localizaton** requirement, making the evaluation process more robust and clinically relevant. Third, the integration of ALPEC into the taxonomy by Sørbø and Ruocco (2023) demonstrates its alignment with established categories in time series anomaly detection. ALPEC metrics classify as *binary*, since thresholding does not manipulate prediction scores, and *redefined counting-based*, as they involve comparing intervals rather than evaluating pointwise. Additionally, they exhibit intrinsic insensitivity to true negatives and a valuation property of time tolerance.

ALPEC also introduces several significant differences: First, ALPEC **merges close predicted intervals**, where we set $\delta = 10s$, in line with clinical guidelines requiring at least 10 seconds of stable sleep between arousals (Berry et al., 2012). Second, ALPEC imposes a **maximum duration** on predicted intervals, which we set to $d = 60s$, ensuring they do not exceed practical lengths. This restriction intends to prevent ambiguities during human review and maintain the clinical relevance of arousal timing relative to sleep stages (that are scored in 30-second epochs) which is important for physicians in sleep medicine. Third, the method of **precise event counting** ensures that only one TP is counted per predicted interval, which is essential for clinical utility. If a predicted interval spans multiple ground-truth intervals, it results in multiple FNs unless each ground-truth interval is uniquely matched to a predicted interval. This approach avoids the pitfall of inaccurately rewarding temporal extension of predicted intervals, a limitation of segment-wise f-score methods (Sørbø and Ruocco, 2023).

Finally, unlike most current approaches to arousal detection, which typically rely on the F1 score or limit their reports to AUPRC or AUROC without a clear consensus on the most appropriate metric (see Table 1), ALPEC follows Sørbø and Ruocco (2023) in advocating for a context-aware selection of metrics

tailored to the operational environment and model selection process. Relying solely on AUPRC or AUROC is inadequate for **clinical decision-making**, as these metrics do not fully capture how model predictions translate into actionable outcomes. For applications, where reliability is crucial, task-specific threshold analysis is preferable, as it directly reflects operational trade-offs. Also, recent research highlights potential biases for AUPRC and questions its applicability in high-stakes decision-making (McDermott et al., 2024). Specifically, AUPRC tends to show high variance in imbalanced datasets with few positive samples, such as arousal events in our case, making it an unreliable criterion for model selection. Therefore, we acknowledge the utility of these metrics in performance analysis but do not consider them sufficient for performance evaluation and model selection. AI-based clinical decision support systems (CDSS) in healthcare aim not only to improve the quality of outcomes but also to enhance the efficiency of medical practitioners' work (Magrabi et al., 2019; Vasey et al., 2022). Given the need to relieve medical practitioners from reviewing extensive amounts of data – in our case, night-long recordings – it is crucial for CDSS to highlight the most pertinent data sections. This objective is best met by ensuring that AI predictions minimize missed arousals (false negatives), allowing human reviewers to efficiently address any false positives. The F2 score is particularly well-suited for this purpose, as it explicitly prioritizes recall over precision, reducing the likelihood of missed arousal events.

3.3. Baseline approaches

In this work, we also explore traditional window-based classification (WBC) methods to provide a comparative baseline for arousal detection. These WBC approaches are evaluated using both standard window-based evaluation and our ALPEC framework. This helps in evaluating the effectiveness of our proposed approach against established techniques.

We employ several classical univariate models from the sktime library (Löning et al., 2019) (BSD 3-Clause License), utilizing them with their standard configurations. For each arousal onset in the training set, we construct a 30-second window centered randomly around the onset point to enhance generalizability. This random alignment aims to mimic the variable alignment of arousal onset points during inference across non-overlapping windows covering the entire series of a subject. To address the challenge of class imbalance, we select an equal number of negative-class windows randomly for each subject during training. The evaluation then proceeds with standard WBC, dividing the test set data into consecutive non-overlapping windows of the same 30-second length used in training. Adopting window-based onset detection simplifies the windowing approach by reducing the need for overlapping windows or complex voting schemes. Instead, we apply the *presence* criterion to determine the class of the windows, aiming to approximate the original class distribution in the test set. Additionally, we ensure robustness by conducting cross-subject validation, where training and testing sets include distinct subjects. Furthermore, we conduct baseline experiments where CS approaches are evaluated using both traditional window-based evaluation and ALPEC for comparative analysis. This involves creating windows for ground-truth and prediction as described above.

4. Results

We perform an experimential comparison of training and evaluation schemes based on the 2018 PhysioNet Challenge Dataset in Section 4.1 and our CPS dataset in Section 4.2. For an illustration of the schemes, refer to Appendix E (Figure 2). Ablation studies on hyperparameter choices are reported in Appendix A.

4.1. Experiments on the 2018 PhysioNet Challenge dataset

The 2018 PhysioNet Challenge Dataset, licensed under the Open Data Commons Attribution License v1.0, is a notable publicly available resource that includes polysomnographic (PSG) data from 1,983 patients at Massachusetts General Hospital's Sleep Lab, with labels provided for 994 subjects (Ghassemi et al., 2018; Goldberger et al., 2000). It adheres to AASM guidelines and includes 13 data channels (six EEG, EOG, EMG at the chin, respiratory at chest and abdomen, ECG, SaO_2, and airflow) annotated with various sleep stages and arousal categories. Annotated events include target arousals (RERA: Respiratory Effort-Related Arousals) and non-target arousals (Hypopnea, Central Apnea, Mixed Apnea, and Obstructive Apnea).

We now explore the shift from full event detection (FED) to interval-based onset detection (IOD) and consider point-based onset detection (POD) as an alternative. We utilize the DeepSleep approach for con-

tinuous segmentation (CS) and evaluate with both the traditional pointwise evaluation (PE) scheme and our ALPEC framework. The 2018 PhysioNet Challenge Dataset provides a basis for training and comparing a full event detection (FED) baseline due to its arousal annotations with both meaningful start and end points. We randomly partitioned the 994 samples into a training set of 795 samples and a test set of 199 samples. Utilizing all 13 channels, training proceeded until either early stopping criteria were met or 50 epochs were completed. The results are shown in Table 2.

Our results indicate that point-based onset detection (POD) using the DeepSleep approach for continuous segmentation is infeasible. Due to the sparsity of labels and noise in the onset annotations, the model is unable to relate meaningful patterns to arousal onsets. In pointwise evaluation (PE), a very low decision threshold results in high recall but low precision, leading to the best possible F2 score, which remains close to zero. ALPEC offers a more realistic assessment of the model's performance by discarding excessively long predicted intervals. However, if a model were to make point-predictions for arousal onsets within the temporal tolerance buffer of ground-truth onset points, ALPEC is expected to perform adequately, whereas pointwise evaluation would penalize predictions that are off by even one point. Thus, ALPEC enables the utilization and appropriate evaluation of other point-based detection approaches that have not been used before for arousal detection, such as methods for changepoint detection in time series (Aminikhanghahi and Cook, 2017).

Moreover, interval-based onset detection (IOD) performs comparably to the FED baseline when measured by ALPEC, demonstrating that detecting arousal onsets rather than full events can be equally effective. Conversely, pointwise evaluation asserts a substantial performance discrepancy, underscoring the importance of choosing an appropriate evaluation framework. An investigation of the large confidence intervals in Table 2 is deferred to Appendix B.

4.2. Experiments on the CPS dataset

The CPS dataset (see also *Data and Code Availability*) comprises 113 diagnostic polysomnographic recordings, encompassing up to 36 raw and 23 derived data channels, alongside 81 types of annotated events and additional questionnaire data for each subject. The dataset annotates various arousal classes, including those related to respiratory efforts, flow limitations, oxygen desaturation, limb movements, and spontaneous arousals. We combine all classes into a single category for binary event detection. Further details on the dataset are available in Appendix H.

Using the CPS dataset, we trained DeepSleep model candidates D1-D4, which utilize continuous segmentation on interval-based onset detection (IOD, see Section 3.1). Details about the selection and channels of the model candidates, as well as the choice of hyperparameters, are provided in Appendix D. These models are compared to popular time series classification models and naive baselines from the sktime library (Löning et al., 2019) using window-based onset detection (WOD, see Section 3.3). All models are trained using combined training and validation folds (93 subjects total). The DeepSleep models are trained until early stopping or up to 100 epochs. All models are evaluated on the held-out test set (14 subjects) using both window-based evaluation (WE) and ALPEC. All experiments are repeated five times using different random seeds. Details on data preprocessing, and data folds are available in Appendices C, and I. The results are shown in Table 3.

Both Window-Based Evaluation (WE) and ALPEC demonstrate similar performance across our interval-based onset (IOD) detection models (D1-D4), which utilize the DeepSleep approach. This consistency suggests that the domain-specific adaptations inherent in ALPEC do not drastically alter results compared to WE in this context. However, WE appears to underestimate precision, resulting in a higher count of false positives. This discrepancy is due to ALPEC's methodology of treating all adjacent points within an overlapping interval as a single true positive, unlike WE, which evaluates each window individually. Furthermore, ALPEC's buffer zones typically reduce false positives and negatives, enhancing its accuracy. In the analysis of Window-Based Onset Detection (WOD) models, WE notably overestimates recall compared to ALPEC. A clear example is observed with the *Constant 1* baseline, where WE significantly overrates its performance because this model predicts an arousal event in every window. This outcome reveals a bias in WE towards models that predict frequent events. Conversely, ALPEC shows zero performance for this baseline, effectively highlighting its ability to address the methodological shortcomings of its closest predecessor, the segment-wise f-score, as discussed in Section 3.2.

Table 2: **Comparison of modeling and evaluation approaches on the 2018 PhysioNet Challenge Dataset**. Arousal types are explained in Section 4.1. Training methods: *FED* (Full Event Detection), *POD* (Point-based Onset Detection), *IOD* (Interval-based Onset Detection) use DeepSleep (Section 3.1). Evaluation: *PE* (Pointwise Evaluation), *ALPEC* (Approximate Localization and Precise Event Count), see Sections 3.3 and 3.2. Metrics are mean values over test subjects with cross-subject validation. Models are trained five times, results averaged with 95% confidence intervals, assuming t-distributed mean values. POD is ineffective with DeepSleep; ALPEC, shows that IOD performs comparable to the FED baseline.

Arousals	Training approach	PE (baseline)			ALPEC evaluation		
		Precision	Recall	F2	Precision	Recall	F2
Target	IOD (our)	0.13 (4)	0.47 (7)	0.30 (3)	0.20 (6)	0.63 (12)	0.42 (4)
	POD (naive baseline)	7e-6	0.94	3.5e-5	0.00	0.00	0.00
	FED (baseline)	0.17 (5)	0.49 (17)	0.35 (8)	0.23 (15)	0.59 (20)	0.41 (6)
Non-target	IOD (our)	0.33 (6)	0.71 (10)	0.57 (3)	0.53 (6)	0.85 (6)	0.76 (3)
	POD (naive baseline)	3.4e-5	1.00	1.67e-4	0.00	0.00	0.00
	FED (baseline)	0.54 (4)	0.67 (8)	0.64 (5)	0.60 (4)	0.77 (5)	0.73 (2)

Table 3: **Comparison of modeling and evaluation approaches on our CPS dataset.** Metrics are presented as mean values over distinct test subjects with 95% confidence intervals, assuming t-distributed mean values, calculated over five training iterations. For window-based onset detection (WOD) and baseline models, methods from the sktime library (Löning et al., 2019), including dummy models, were used. Models D1-D4, derived from the DeepSleep approach utilizing interval-based onset detection (IOD), indicate the used channels. ALPEC offers a stringent assessment of performance, highlighting that WE tends to overestimate the effectiveness of WOD models.

	Model	Window-based evaluation (WE)			ALPEC evaluation		
		Precision	Recall	F2	Precision	Recall	F2
IOD	D4: most channels	0.49 (6)	0.82 (8)	**0.71** (3)	0.59 (8)	0.81 (8)	**0.73** (3)
	D3: no EEG, EOG, EMG	0.39 (3)	0.71 (3)	0.59 (3)	0.48 (4)	0.70 (3)	0.62 (3)
	D2: C3:A2, EOGl, EMG	0.44 (5)	0.75 (7)	0.64 (3)	0.53 (7)	0.74 (7)	0.67 (3)
	D1: C3:A2	0.40 (5)	0.76 (6)	0.62 (2)	0.48 (7)	0.75 (6)	0.65 (2)
WOD	IndividualBOSS	0.25 (0)	0.55 (1)	0.42 (1)	0.30 (1)	0.32 (2)	0.31 (2)
	SupervisedTimeSeriesForest	0.37 (0)	0.71 (1)	0.59 (1)	0.37 (1)	0.29 (1)	0.30 (1)
	TimeSeriesForestClassifier	0.30 (1)	0.65 (1)	0.50 (1)	0.30 (1)	0.30 (1)	0.29 (1)
	SignatureClassifier	0.28 (0)	0.65 (2)	0.49 (1)	0.29 (1)	0.30 (1)	0.29 (1)
	SummaryClassifier	0.27 (0)	0.62 (1)	0.48 (0)	0.27 (1)	0.29 (1)	0.28 (1)
	Catch22Classifier	0.33 (0)	0.73 (1)	0.57 (0)	0.30 (0)	0.23 (1)	0.24 (1)
Baseline	RandomStratified	0.20 (1)	0.49 (2)	0.37 (1)	0.29 (2)	0.36 (4)	0.34 (3)
	RandomUniform	0.20 (0)	0.51 (2)	0.37 (1)	0.28 (2)	0.36 (3)	0.33 (3)
	Constant 1	0.20	1.00	0.53	0.00	0.00	0.00
	Constant 0	0.00	0.00	0.00	0.00	0.00	0.00

Overall, ALPEC provides a more accurate assessment of model performance, revealing that none of the WOD models substantially outperform the random baselines. This aligns with the understanding that arousal detection is a challenging task that may not be adequately addressed by simpler classical mod-

els without specialized feature engineering (Zan and Yildiz, 2023).

Finally, we find a significantly enhanced predictive performance of model D4, which incorporates the most data modalities, compared to models D1 and D2, which use fewer modalities. Also, model D3, which does not use any electrode-based modalities, demonstrates potential for reduced technical complexity while maintaining reasonable performance.

5. Discussion

Our findings demonstrate that arousal onset detection can be effectively achieved using continuous segmentation approaches with our proposed interval-based onset detection (IOD) training scheme, achieving comparable performance to the full event detection (FED) baseline, successfully aligning model training with clinical annotation constraints for arousal annotations.

Additionally, the results highlight the significant benefits of incorporating novel data modalities, which also offer potential for reducing technical complexity. Minimizing dependence on electrode-based modalities could address issues such as electrode displacement or noise, potentially enabling home-based arousal diagnostics (Imtiaz, 2021).

A significant contribution of our work is the development of the ALPEC framework, the first performance evaluation framework tailored to the clinical requirements of arousal detection. We demonstrate that ALPEC provides a more accurate assessment of model performance compared to traditional window-based evaluation (WE) or pointwise evaluation (PE). Moreover, due to sampling at the subject-level, ALPEC overcomes common pitfalls of window-based evaluation such as class imbalance and cross-subject validation issues (see Appendix G for further details). We emphasize that our critique is not directed against window-based *classification* approaches, which remain valuable and effective in arousal detection, as shown in recent studies (Badiei et al., 2023; Foroughi et al., 2023). Our concerns specifically relate to window-based *evaluation* methods.

We advocate for the adoption of the ALPEC framework, which is immune to common pitfalls, finely tunable, and compatible with both window-based classification and continuous segmentation. ALPEC's design inherently incorporates a *precise event count* requirement while offering flexibility in *approximate location* through adjustable parameters for buffer size and maximum interval length. Additionally, it is suitable to evaluate models trained on various ground-truth annotations, including point annotations (POD), constructed intervals (IOD), and events with start and end values (FED). This adaptability makes ALPEC a versatile tool for any task requiring precise event count detection in time series data, providing a robust framework suitable for a wide range of applications in healthcare and beyond.

5.1. Limitations

(1) This work addresses binary detection of arousal onsets and does not encompass the causal differentiation of arousals, an additional task in clinical settings that necessitates a multi-class classification approach.

(2) Furthermore, ALPEC is designed solely for post-processing and performance evaluation, and does not influence the learning process of the models. While we adapted the DeepSleep method for arousal onset detection, it was not initially designed to meet the specific demands of real-world clinical applications. Future research should leverage ALPEC for comparative analysis and enhance model functionality by incorporating factors crucial for clinical decision support, such as explainability.

(3) Additionally, the settings for ALPEC's buffer size and maximum interval length hyperparameters need experimental validation in collaboration with clinical end-users to ensure their effectiveness and applicability in real-world settings.

We will address limitations two and three through an application-grounded user study in future work.

5.2. Conclusion

Our work establishes foundational elements for developing clinical decision support systems for arousal detection in sleep laboratories, addressing critical misalignments between current Machine Learning methodologies and clinical practices. We introduce the Comprehensive Polysomnography (CPS) dataset as a significant resource for sleep medical research, demonstrating the potential of utilizing novel data modalities.

Our findings contribute to the development of production-ready arousal detection models that align with current clinical annotation practices. We look forward to seeing how the research community builds on our findings and continues to evolve the field.

References

Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Joan Giner-Miguelez, Nitisha Jain, Michael Kuchnik, Quentin Lhoest, Pierre Marcenac, Manil Maskey, Peter Mattson, Luis Oala, Pierre Ruyssen, Rajat Shinde, Elena Simperl, Goeffry Thomas, Slava Tykhonov, Joaquin Vanschoren, Steffen Vogler, and Carole-Jean Wu. Croissant: A Metadata Format for ML-Ready Datasets, 2024.

Samaneh Aminikhanghahi and Diane J Cook. A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367, 2017.

Adriana Anido-Alonso and Diego Alvarez-Estevez. Decentralized data-privacy preserving deep-learning approaches for enhancing inter-database generalization in automatic sleep staging. *IEEE Journal of Biomedical and Health Informatics*, 2023.

Jerome Argod, Jean-Louis Pepin, and Patrick Levy. Differentiating obstructive and central sleep respiratory events through pulse transit time. *American journal of respiratory and critical care medicine*, 158(6):1778–1783, 1998.

Afsoon Badiei, Saeed Meshgini, and Khosro Rezaee. A novel approach for sleep arousal disorder detection based on the interaction of physiological signals and metaheuristic learning. *Computational Intelligence and Neuroscience*, 2023, 2023.

Richard B Berry, Rohit Budhiraja, Daniel J Gottlieb, David Gozal, Conrad Iber, Vishesh K Kapur, Carole L Marcus, Reena Mehra, Sairam Parthasarathy, Stuart F Quan, et al. Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events: deliberations of the sleep apnea definitions task force of the American Academy of Sleep Medicine. *Journal of clinical sleep medicine*, 8(5):597–619, 2012.

Bertrand. SweetViz. Visualize and compare datasets, target values and associations, with one line of code. https://github.com/fbdesignpro/sweetviz, 2020. Accessed: 2024-06-04.

Maria R Bonsignore, Tarja Saaresranta, and Renata L Riha. Sex differences in obstructive sleep apnoea. *European Respiratory Review*, 28(154), 2019.

Xiaoli Chen, Rui Wang, Phyllis Zee, Pamela L Lutsey, Sogol Javaheri, Carmela Alcántara, Chandra L Jackson, Michelle A Williams, and Susan Redline. Racial/ethnic differences in sleep disturbances: the Multi-Ethnic Study of Atherosclerosis (MESA). *Sleep*, 38(6):877–888, 2015.

Franz Ehrlich, Tony Sehr, Moritz Brandt, Martin Schmidt, Hagen Malberg, Martin Sedlmayr, and Miriam Goldammer. State-of-the-art sleep arousal detection evaluated on a comprehensive clinical dataset. *Scientific Reports*, 14(1):16239, 2024.

Ahmad Fawzy, Danastri Cantya Nirmala, Denaya Khansa, and Yudhistira Tri Wardhana. Ethics and Regulation for Artificial Intelligence in Healthcare: Empowering Clinicians to Ensure Equitable and High-Quality Care. *International Journal of Medical Science and Clinical Research Studies*, 3(07):1350–1357, 2023.

Ingo Fietze, Naima Laharnar, Anne Obst, Ralf Ewert, Stephan B Felix, Carmen Garcia, Sven Gläser, Martin Glos, Carsten Oliver Schmidt, Beate Stubbe, et al. Prevalence and association analysis of obstructive sleep apnea with gender and age differences–Results of SHIP-Trend. *Journal of sleep research*, 28(5):e12770, 2019.

Robert Fonod. DeepSleep 2.0: automated sleep arousal segmentation via deep learning. *AI*, 3(1):164–179, 2022.

Andia Foroughi, Fardad Farokhi, Fereidoun Nowshiravan Rahatabad, and Alireza Kashaninia. Deep convolutional architecture-based hybrid learning for sleep arousal events detection through single-lead EEG signals. *Brain and Behavior*, 13(6):e3028, 2023.

Karl A Franklin and Eva Lindberg. Obstructive sleep apnea is a common disorder in the population—a review on the epidemiology of sleep apnea. *Journal of thoracic disease*, 7(8):1311, 2015.

Azul Garza and Max Mergenthaler-Canseco. TimeGPT-1. *arXiv preprint arXiv:2310.03589*, 2023.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Mohammad M Ghassemi, Benjamin E Moody, Li-Wei H Lehman, Christopher Song, Qiao Li, Haoqi Sun, Roger G Mark, M Brandon Westover, and Gari D Clifford. You snooze, you win: the physionet/computing in cardiology challenge 2018. In *2018 Computing in Cardiology Conference (CinC)*, volume 45, pages 1–4. IEEE, 2018.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation*, 101(23): e215–e220, 2000.

Matthew Howe-Patterson, Bahareh Pourbabaee, and Frederic Benard. Automated detection of sleep arousals from polysomnography data using a dense convolutional neural network. In *2018 Computing in Cardiology Conference (CinC)*, volume 45, pages 1–4. IEEE, 2018.

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 387–395, 2018.

Syed Anas Imtiaz. A systematic review of sensing technologies for wearable sleep staging. *Sensors*, 21(5): 1562, 2021.

Shazia Jehan, Ferdinand Zizi, Seithikurippu R Pandi-Perumal, Steven Wall, Evan Auguste, Alyson K Myers, Girardin Jean-Louis, and Samy I McFarlane. Obstructive sleep apnea and obesity: implications for public health. *Sleep medicine and disorders: international journal*, 1(4), 2017.

Stefan Kraft, Andreas Theissler, Vera Wienhausen-Wilke, Philipp Walter, and Gjergji Kasneci. Comprehensive Polysomnography (CPS) Dataset: A Resource for Sleep-Related Arousal Research. PhysioNet data repository, 2024. URL https://doi.org/10.13026/sxs0-h317. Dataset.

Chih-Fan Kuo, Cheng-Yu Tsai, Wun-Hao Cheng, Wen-Hua Hs, Arnab Majumdar, Marc Stettler, Kang-Yun Lee, Yi-Chun Kuan, Po-Hao Feng, Chien-Hua Tseng, et al. Machine learning approaches for predicting sleep arousal response based on heart rate variability, oxygen saturation, and body profiles. *Digital Health*, 9: 20552076231205744, 2023.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the Carbon Emissions of Machine Learning. *arXiv preprint arXiv:1910.09700*, 2019.

H Lee, B Li, Y Huang, Y Chi, and S Lin. NCH sleep databank: a large collection of real-world pediatric sleep studies with longitudinal clinical data (version 3.1. 0). PhysioNet, 2021.

Harlin Lee, Boyue Li, Shelly DeForte, Mark L Splaingard, Yungui Huang, Yuejie Chi, and Simon L Linwood. A large collection of real-world pediatric sleep studies. *Scientific Data*, 9(1):421, 2022.

Haoqi Li, Qineng Cao, Yizhou Zhong, and Yun Pan. Sleep arousal detection using end-to-end deep learning method based on multi-physiological signals. In *2018 computing in cardiology conference (CinC)*, volume 45, pages 1–4. IEEE, 2018.

Hongyang Li and Yuanfang Guan. DeepSleep convolutional neural network allows accurate and fast detection of sleep arousal. *Communications biology*, 4(1):18, 2021.

Markus Löning, Anthony Bagnall, Sajaysurya Ganesh, Viktor Kazakov, Jason Lines, and Franz J Király. sktime: A unified interface for machine learning with time series. *arXiv preprint arXiv:1909.07872*, 2019.

Tambiama Madiega. Artificial intelligence act. *European Parliament: European Parliamentary Research Service*, 2021.

Farah Magrabi, Elske Ammenwerth, Jytte Brender McNair, Nicolet F De Keizer, Hannele Hyppönen, Pirkko Nykänen, Michael Rigby, Philip J Scott, Tuulikki Vehko, Zoie Shui-Yee Wong, et al. Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implications. *Yearbook of medical informatics*, 28(01):128–134, 2019.

Matthew McDermott, Lasse Hyldig Hansen, Haoran Zhang, Giovanni Angelotti, and Jack Gallifant. A Closer Look at AUROC and AUPRC under Class Imbalance. *arXiv preprint arXiv:2401.06091*, 2024.

Daniel Miller, Andrew Ward, and Nicholas Bambos. Automatic sleep arousal identification from physiological waveforms using deep learning. In *2018 Computing in Cardiology Conference (CinC)*, volume 45, pages 1–4. IEEE, 2018.

Tomofumi Misaka, Yuko Niimura, Akiomi Yoshihisa, Kento Wada, Yusuke Kimishima, Tetsuro Yokokawa, Satoshi Abe, Masayoshi Oikawa, Takashi Kaneshiro, Atsushi Kobayashi, et al. Clinical impact of sleep-disordered breathing on very short-term blood pressure variability determined by pulse transit time. *Journal of Hypertension*, 38(9):1703–1711, 2020.

Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27 (3):400–410, 2019.

D Pitson, N Chhina, S Knijn, M Van Herwaaden, and J Stradling. Changes in pulse transit time and pulse rate as markers of arousal from sleep in normal subjects. *Clinical science (London, England: 1979)*, 87(2): 269–273, 1994.

DJ Pitson et al. Value of beat-to-beat blood pressure changes, detected by pulse transit time, in the management of the obstructive sleep apnoea/hypopnoea syndrome. *European Respiratory Journal*, 12(3):685–692, 1998.

Naresh M Punjabi. The epidemiology of adult obstructive sleep apnea. *Proceedings of the American Thoracic Society*, 5(2):136–143, 2008.

Stuart F Quan, Barbara V Howard, Conrad Iber, James P Kiley, F Javier Nieto, George T O'Connor, David M Rapoport, Susan Redline, John Robbins, Jonathan M Samet, et al. The sleep heart health study: design, rationale, and methods. *Sleep*, 20(12):1077–1085, 1997.

Winfried Randerath, Claudio L Bassetti, Maria R Bonsignore, Ramon Farre, Luigi Ferini-Strambi, Ludger Grote, Jan Hedner, Malcolm Kohler, Miguel-Angel Martinez-Garcia, Stefan Mihaicuta, et al. Challenges and perspectives in obstructive sleep apnoea: report by an ad hoc working group of the Sleep Disordered Breathing Group of the European Respiratory Society and the European Sleep Research Society. *European respiratory journal*, 52(3), 2018.

F Raschke and J Fischer. "Arousal" in der Schlafmedizin. *Somnologie*, 1(2), 1997.

Erik Scharwächter and Emmanuel Müller. Statistical evaluation of anomaly detectors for sequences. *arXiv preprint arXiv:2008.05788*, 2020.

Daniel J Schwartz and Pat Moxley. On the potential clinical relevance of the length of arousals from sleep in patients with obstructive sleep apnea. *Journal of Clinical Sleep Medicine: JCSM: Official Publication of the American Academy of Sleep Medicine*, 2(2):175–180, 2006.

Sobhan Salari Shahrbabaki, Dominik Linz, Simon Hartmann, Susan Redline, and Mathias Baumert. Sleep arousal burden is associated with long-term all-cause and cardiovascular mortality in 8001 community-dwelling older men and women. *European heart journal*, 42(21):2088–2099, 2021.

Sondre Sørbø and Massimiliano Ruocco. Navigating the metric maze: A taxonomy of evaluation metrics for anomaly detection in time series. *Data Mining and Knowledge Discovery*, pages 1–42, 2023.

Baptiste Vasey, Myura Nagendran, Bruce Campbell, David A Clifton, Gary S Collins, Spiros Denaxas, Alastair K Denniston, Livia Faes, Bart Geerts, Mudathir Ibrahim, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *bmj*, 377, 2022.

Thomas-Christian Wetter, Roland Popp, Michael Arzt, and Thomas Pollmächer. *ELSEVIER ESSENTIALS Schlafmedizin: Das Wichtigste für Ärzte aller Fachrichtungen*. Elsevier Health Sciences, 2012.

Wienhausen-Wilke and Kraft. Computer-aided diagnostics of sleep-related arousals on the basis of pulse wave analyses. https://drks.de/search/en/trial/DRKS00033641, 2024. [Accessed: 2024-08-15].

Hasan Zan and Abdulnasır Yildiz. Multi-task learning for arousal and sleep stage detection using fully convolutional networks. *Journal of Neural Engineering*, 20(5): 056034, 2023.

Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The National Sleep Research Resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, 25(10):1351–1358, 2018.

Appendix A. Ablation studies

In this section, we present ablation studies conducted on our the CPS dataset to evaluate the impact of various hyperparameters on the performance of the DeepSleep method for arousal onset detection and the ALPEC framework for performance evaluation. The hyperparameters considered are parameters used in training (the smoothing window size w, and the interval length for interval-based onset detection (IOD) l) and the ALPEC framework (the maximum interval duration d, the minimum interval distance before merging δ, and the buffer size b). For a description of the parameters, see Table 9. We use the DeepSleep model architecture D1 from Table 3 with a univariate *C3:A2* channel. All models are trained using the same training and test split as in Section 4. Tables 4 to 8 show the tuning results, where parameter choices for all runs, if not tuned, are marked in bold, which are the same as for D1 in the main part of this paper.

Table 4: Smoothing window w

w	$\bar{\text{Precision}}$	$\bar{\text{Recall}}$	$\bar{\text{F2}}$
none	0.47	0.67	0.60
1	0.59	0.71	0.66
2	0.53	0.73	0.65
3	0.47	0.80	0.68
4	0.45	0.86	0.70
5	0.44	0.82	0.67

Table 5: Interval length for IOD l

l	$\bar{\text{Precision}}$	$\bar{\text{Recall}}$	$\bar{\text{F2}}$
2	0.39	0.78	0.62
6	0.49	0.67	0.61
10	0.45	0.80	0.68
14	0.47	0.80	0.68
20	0.42	0.81	0.66
30	0.47	0.59	0.55
60	0.36	0.50	0.45

We can see that most parameters have a moderate effect on performance metrics within the ranges tested. The most significant drop in performance occurs with low values of the maximum allowed interval distance d, which is expected since removing many events leads to a high number of false negatives. Values higher than $d = 60s$ make no significant difference to $d = 60s$, indicating that our ML approach produces reasonably short predicted intervals. We also see that smoothing (Table 4), merging of intervals (Table 7), and utilizing a buffer (Table 8) all lead to performance improvements. Smoothing and merging actually affect the predicted intervals, whereas the buffer only affects the evaluation by relaxing the locality requirement.

Table 6: Max. interval duration d

d	$\bar{\text{Precision}}$	$\bar{\text{Recall}}$	$\bar{\text{F2}}$
10	0.57	0.19	0.22
30	0.50	0.74	0.66
60	0.47	0.80	0.68
90	0.43	0.77	0.65
120	0.40	0.86	0.67
none	0.45	0.82	0.68

Table 7: Min. interval distance before merging δ

δ	$\bar{\text{Precision}}$	$\bar{\text{Recall}}$	$\bar{\text{F2}}$
0	0.48	0.76	0.66
5	0.50	0.74	0.66
10	0.47	0.80	0.68
15	0.37	0.86	0.66
20	0.37	0.58	0.50

Table 8: Buffer size b with $b = b^{\text{before}} = b^{\text{after}}$

b	$\bar{\text{Precision}}$	$\bar{\text{Recall}}$	$\bar{\text{F2}}$
0	0.51	0.69	0.63
5	0.47	0.70	0.61
10	0.46	0.75	0.65
15	0.47	0.80	0.68
20	0.43	0.82	0.68
25	0.59	0.66	0.63

Appendix B. Analysis of the effects of decision thresholds

The rather large confidence intervals in the results on the 2018 PhysioNet Challenge Dataset (Section 4,

Table 9: Table of notation

Symbol	Meaning
$\mathbf{x}$	Multivariate input sequence
n	Number of data points contained within each input channel after padding with zeros, fixed to 2^{23}
D	Dataset containing all subjects
T	Training set containing a subset of subjects
V	Validation set containing a subset of subjects
$p_i(\mathbf{x})$	Probability score of the i-th time step in the input sequence being of the positive class
$p_\eta(\mathbf{x})$	Probability score for the η-th window in the input sequence
$c_i(\mathbf{x})$	Binary class prediction for the i-th time step
N	Number of windows when splitting the input sequence into windows of length s
s	Length of each window when splitting the input sequence into N windows
w	Window size for smoothing the probability scores
f	Sampling frequency of the data
t_k	Threshold for converting probability scores to binary class predictions. We use 101 thresholds from 0 to 1 in steps of 0.01, i.e. $k = 0, 1, ..., 100$
δ	Minimum distance in seconds for merging two adjacent predicted intervals in ALPEC
C	Binary class predictions for the whole input sequence, i.e. $C = \{c_1, c_2, ..., c_n\}$
I	Predicted interval in binary class predictions C with start and end indices I^{start} and I^{end}
G	Ground-truth interval with start and end indices G^{start} and G^{end}
d	Maximum duration of a predicted interval before its removal in ALPEC
P	Predicted interval with start and end indices P^{start} and P^{end}
b^{before}	Temporal tolerance buffer before the ground-truth interval in ALPEC
b^{after}	Temporal tolerance buffer after the ground-truth interval in ALPEC
G^{ext}	Extended ground-truth interval with start and end indices $G^{\text{start,ext}}$ and $G^{\text{end,ext}}$
l	Length of the interval around an onset point for interval-based onset detection (IOD)

Table 2) stem from the variance in selected decision thresholds for individual runs, as shown in Table 10.

Table 10: **Comparison of selected decision thresholds (DT) and their effects**. This table presents the decision thresholds for five individual runs with different random seeds, leading to the results for the *Target* arousals using *FED* training and *ALPEC* evaluation shown in Table 2.

Run	DT	$\overline{\text{Precision}}$	$\overline{\text{Recall}}$	$\overline{\text{F1}}$	$\overline{\text{F2}}$
1	0.22	0.44	0.42	0.43	0.43
2	0.03	0.17	0.82	0.47	0.47
3	0.07	0.18	0.68	0.44	0.44
4	0.13	0.19	0.48	0.37	0.37
5	0.12	0.15	0.53	0.35	0.35

These thresholds are based on samples from the training fold and are automatically selected to maximize the F2 score. Comparing runs 1 and 2, for example, shows that a lower decision threshold results in higher recall but lower precision, as expected, while the resulting F2 scores are comparable.

Appendix C. Data preprocessing

For our experiments, all raw data channels undergo third-order Butterworth bandpass filtering to remove noise. Critical frequencies for the Butterworth bandpass filter for the different data modalities are listed in Table 11.

The raw data channels are then normalized using z-score normalization. Derived channels are upsampled to 256 Hz using repeated values and scaled to a range of $[0, 1]$ via min-max normalization. Channels are padded symmetrically to a fixed length of $n = 2^{23}$, or approximately 9 hours, to accommodate the longest recording. Magnitude scaling is applied randomly between 0.8 and 1.25 during training to enhance model generalization Li and Guan (2021).

All nominal event data are encoded as binary features, with each event type represented as a separate feature. For Sleep Profile and Body Position events, we utilize a one-hot encoding scheme to represent the different classes.

Appendix D. Hyperparameter tuning details

In this section, we provide an overview of the selected hyperparameters and perform preliminary experiments with the DeepSleep approach for continuous segmentation on unimodal data channels to find a good set of input channels for final model candidates. All models are trained on the training set and evaluated on a fixed validation set. See Appendix I for details on the data splits. Table 12 contains an overview of the hyperparameters used in this work.

Selection of input channels We perform two baseline sets of tuning runs: One on raw channels (see Table 15) and another on derived channels and a promising selection of event channels (see Tables 16 and 17). From the raw channels, we leave out the *Battery* and *REM Confidence* channels since we expect those to be irrelevant for arousal detection. Also, we only use one channel each from the EEG, EMG, and EOG groups. We split the categorical event channels *Sleep Profile* and *Body Position* into singular channels using a one-hot encoded representation of the categories. For simplicity, we will refer to both the derived and event channels as *derived* channels from here on. All results from these runs are shown in Tables 13 and 14.

Tuning all possible combinations of raw and derived channels from Tables 15, 16, and 17 would be computationally very demanding, even when restricting ourselves to the most discriminative channels. From explorative experiments, we learned that the performance of DeepSleep generally increases when using additional channels. Therefore, we selected four combinations of channels as model candidates for the final evaluation in the main part of this work, denoted with a *D* for *DeepSleep*:

1. D1, using only the channel *C3:A2*, which yielded the best performance in Table 13 and is the required choice for manual arousal detection according to the AASM guidelines (Berry et al., 2012).

2. D2, using modalities often selected for arousal detection in related work (see Table 1), namely *C3:A2*, *EOGl*, and *EMG*.

3. D3, using a selection of channels that do not rely on EEG, EMG, and EOG modalities as indicated in Tables 13 and 14 in the *D3* column.

Table 11: Critical frequencies for the bandpass filter for different modalities

Modality	**Channels**	**Lower freq. [Hz]**	**Upper freq. [Hz]**
EEG	C4:A1, C3:A2, F4:A1, O2:A1, A1, A2, C3, C4, F4, O2	0.2	35
EOG	EOGl, EOGL:A1, EOGL:A2, EOGr, EOGr:A1, EOGr:A2	0.2	35
EMG	EMG+, EMG-, EMG	10	127
ECG	ECG 2	0.2	127
Respiratory	Pressure Flow, Thermal Flow	0.001	15
Snore	Snoring Pressure, Snoring Sound	20	127
PPG	Pleth	0.5	5

Table 12: **Choices for Hyperparameters**. Values in seconds are multiplied by the fixed sampling rate of 256 Hz. For further explanations of the meaning of the symbols, see Table 9

Context	Parameter	Value	Explanation
Data (CPS)	$\|T\|$	64	Number of subjects in the training set
	$\|V\|$	28	Number of subjects in the validation set
	$\|E\|$	14	Number of subjects in the test set
Training	n	2^{23}	Number of padded data points per channel
	s	$30s$	Window size for window-based classification
	ω	3s	Smoothing window for continuous segmentation
	l	10s	Interval length for IOD
	epochs	100	Maximum number of training epochs
	batch size	1	Number of subjects per batch
ALPEC	d	60s	Maximum interval duration
	δ	10s	Minimum interval distance
	b^{before}	15s	Left buffer size
	b^{after}	15s	Right buffer size

Table 13: Unimodal training on raw data channels. *D3* and *D4* are selections of channels that are used as model candidates in the main part of this paper.

Channel	$\bar{F}_2$	D3	D4
C3:A2	0.60		✓
Pressure Flow	0.56	✓	✓
RIP.Abdom	0.53	✓	✓
EMG	0.53		✓
Sum RIPs	0.52	✓	✓
EOGl	0.50		✓
Pulse	0.48	✓	✓
RIP.Thrx	0.46	✓	✓
Pleth	0.45	✓	✓
Snoring Pressure	0.45	✓	✓
Thermal Flow	0.43	✓	✓
PLMl	0.37	✓	✓
ECG 2	0.36	✓	✓
Light	0.34	✓	✓
SPO2	0.33	✓	✓
Snoring Sound	0.30	✓	✓
Motion	0.28	✓	✓

4. D4, using all channels from Tables 13 and 14 except for the most underperforming channels, *Body Position* and *Central Apnea*, also indicated in the *D4* column of the tables.

Additionally, we did not select the *Heart Rate*, *Light*, and *SpO2* derived channels for model candidates *D3* and *D4* since these are very similar to the *Pulse*, *Light*, and *SPO2* raw channels, respectively, as indicated by the similar performances in Tables 13 and 14.

The effects of additional hyperparameters are detailed in Appendix A, where calculations are performed using the *D1* unimodal channel selection.

Appendix E. Schematic comparison of training and evaluation schemes

In this work, we utilize a multitude of training and evaluation approaches which are schematically illustrated in Figure 2. For an explanation of the schemes, we refer to Section 3 and Appendix F.

We now want to perform a more detailed conceptual comparison of our proposed Approximate Lo-

Table 14: Unimodal training on derived channels. *D3* and *D4* are selections of channels that are used as model candidates in the main part of this paper.

Channel	$\bar{F}_2$	D3	D4
Average Frequency Value	0.55		✓
Hypopnea	0.53	✓	✓
Sigma FFT	0.50		✓
Heart Rate	0.48		
RR Interval	0.45	✓	✓
Delta FFT	0.45		✓
Alpha+Beta FFT	0.44		✓
PTT Raw	0.44	✓	✓
HRV LF	0.43	✓	✓
Diastol	0.43	✓	✓
Obstruction	0.43	✓	✓
Systol PTT	0.42	✓	✓
Sleep Profile	0.42		✓
Syst	0.42	✓	✓
Diastol PTT	0.41	✓	✓
RR	0.41	✓	✓
SVB	0.41	✓	✓
Phase Angle	0.41	✓	✓
Light	0.33		
SpO2	0.31		
Activity	0.28	✓	✓
Integral EMG	0.26		✓
HRV HF	0.25	✓	✓
Obstructive Apnea	0.17	✓	✓
Apnea	0.09	✓	✓
Body Position	0.04		
Central Apnea	0.04		

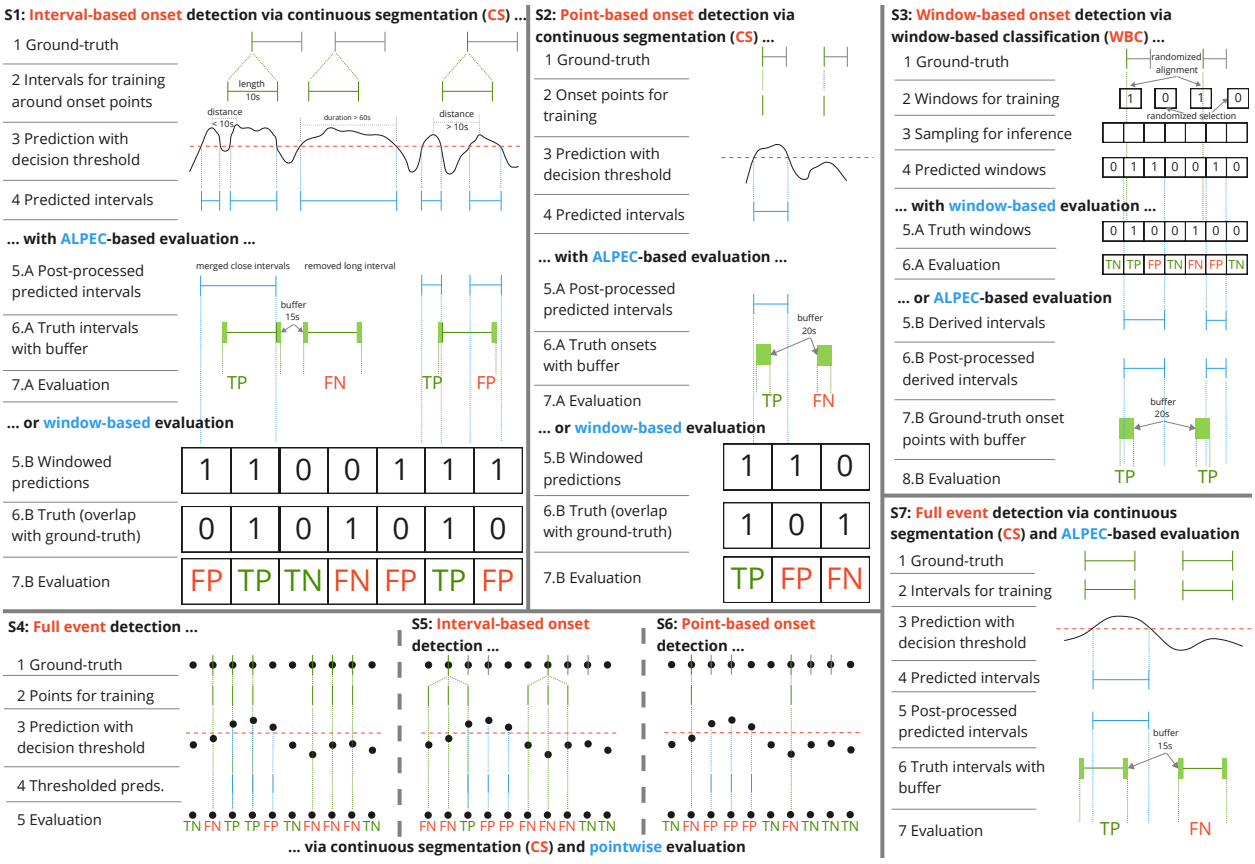

Figure 2: **Schematic illustration of different approaches for training and evaluating arousal detection models.** In schemas S1, S2, S3, and S7, lines and areas in green color represent target points of the positive class (arousal) while empty areas in between contain points of the negative class (no arousal). Lines in blue color represent points that are predicted to be in the positive class. For schemas containing pointwise evaluations (S4-S6), all points which are marked in green or blue are considered to be in the positive class while all other points are considered to be in the negative class. Names of training schemes are highlighted in red, evaluation schemes in blue. All sizes and dimensions are for illustrative purposes and not representative. Especially schemas containing pointwise evaluations will contain many more data points inside events/intervals (S4-S5) and between events (S4-S6). For schemas containing window-based approaches (S1-S3), each box represents a window of fixed length containing many data points, where the class identification or evaluation outcome of each point is given by the label on the box.

calization and Precise Event Count (ALPEC) framework with the baselines of pointwise evaluation and window-based evaluation. We start by remembering that, from a clinical productional point-of-view, the most important aspect of arousal diagnostics is to detect the correct number of arousals and to locate them approximately correctly to enable human validation (cf. Section 3.2). Looking at the *Evaluation* steps of S4-S6 in Figure 2, we see that pointwise evaluation is inadequate based on our requirements since it sanctions every wrong prediction point. This leads to a bias towards favoring models that strictly predict the exact labeled points which might also result in overfitting and a lack of the generalization capabilities of models that are optimized under pointwise evaluation. When comparing the exact situation with ALPEC instead of pointwise evaluation (S7), we see that ALPEC is not concerned with single points but only close-by intervals and counts exactly one TP and one FN as would be expected in this situation from a clinical point-of-view. As Sørbø and Ruocco (2023) have noted generally about pointwise evaluation, a major shortcoming is the lack of tolerance which renders it inappropriate for the evaluation of arousal detection models.

Moving on to window-based evaluation, we find a similar situation as with pointwise evaluation although less severe. Looking at S3, we see that window-based evaluation sanctions the third predicted window although adjacent to a correct prediction (the second) and sanctions the second arousal twice (one FN, one FP), for the close miss of the onset point at the border to the window with a predicted 1. To be fair, window-based evaluation could be equipped with similar domain-specific adaptations like the interval-based ALPEC. Our fundamental critique, however, is that with a windowing approach there are always technical constraints due to the window size which lead to deviations from the intended goal which can be utterly avoided by using the interval-based approach of ALPEC. It entails intrinsic flexibility, allowing it to be closely adapted to the clinical needs. Apart from this, as we have seen in Section 2.1, window-based classification approaches often contain many hyperparameters related to window size, overlap, and voting strategies which often extend to window-based evaluation. ALPEC contains hyperparameters of its own which, however, are less technical and instead are introduced to foster an adaptation to the productive clinical requirements.

Appendix F. Algorithmic description of ALPEC

Algorithm 1 provides an algorithmic description of ALPEC.

ALPEC is compatible with both window-based classification (WBC) and continuous segmentation (CS) approaches to arousal detection. For WBC, we process either probability scores $p_\eta(\mathbf{x})$ or binary predictions c_η for each window η. For CS, the process begins with probability scores $p_i(\mathbf{x})$ for each data point i. In cases where probability scores are available, the optimal threshold t_{opt} is determined from the training set T to convert these scores into binary predictions (Algorithm 1, line 2). For each subject ν and threshold t, we post-process the predictions by applying thresholding and resampling for WBC (lines 17 and 18) or just thresholding for CS (line 20). Next, predicted intervals that are less than $\delta = 10$ seconds apart are merged (line 21). For full event detection, merging is based on the closest points of predicted intervals, while for arousal onset detection, it is based on the maxima of the prediction scores, indicating the most likely points of arousal onset.

After post-processing, predictions are compared to ground-truth data G – which may consist of full event annotations, point annotations, or constructed intervals (see Section 3.1) – to determine true positive (TP), false positive (FP), and false negative (FN) counts (line 12).

ALPEC introduces two key *approximate localization* components. First, a temporal tolerance buffer (Scharwächter and Müller, 2020) of 15 seconds is applied before (b^{before}) and after (b^{after}) each ground-truth interval (line 25). Predicted intervals overlapping with ground-truth intervals within this buffer are counted as TPs. Second, ALPEC restricts the maximum duration of predicted intervals, with only those shorter than $d = 60$ seconds qualifying as TPs (line 27).

The counting method in ALPEC (lines 26 to 32) fulfills the *precise event count* requirement. A TP is recorded when any eligible predicted interval P overlaps with a buffered ground-truth interval G. A FN is recorded if a G does not overlap with any P, and a FP is noted if a P does not overlap with any G. Overly long predicted intervals contribute to only one TP, and multiple ground-truth intervals spanned by a single predicted interval result in multiple FNs unless each ground-truth interval is uniquely matched to a predicted interval.

Algorithm 1 ALPEC post-processing and performance evaluation framework. This compact representation assumes data and main input as globally accessible. The *Eval* function is a placeholder for known implementations in the literature to calculate metrics from TP, FP and FN counts. For-loop variables used outside their scope imply storage in accumulative data structures.

Data: Multivariate input channels $\mathbf{x}$, training set T, validation set V, ground-truth intervals G

Input: Probability scores $p_i(\mathbf{x})$ or $p_\eta(\mathbf{x})$ for each data point i or window η or binary predictions c_η for each window, and hyperparameters: Minimum interval merge distance δ, maximum predicted interval duration d, ground-truth temporal tolerance buffers b^{before} and b^{after}

Output: Mean values for precision, recall and F2-score over subjects in V

1 **if** *Input contains probability scores* $p_i(\mathbf{x})$ *or* $p_\eta(\mathbf{x})$ **then**
2 $t_{\text{opt}} \leftarrow$ `DetermineOptimalThresholdOnTrainingSet()` ; `// Get optimal threshold`
3 **else** `//` Input contains binary predictions c_η
4 $t_{\text{opt}} \leftarrow None$; `// No thresholding`
5 **foreach** *subject* ν *in* V **do**
6 $\text{precision}_\nu, \text{recall}_\nu, \text{F2}_\nu \leftarrow$ `Eval(CompareTruthPredPerSubject(PostProcPreds(`ν, t_{opt}`)))`
7 $\bar{\text{precision}}, \bar{\text{recall}}, \bar{\text{F2}} \leftarrow$ Compute mean values over subjects ν ; `// Get` $\bar{\text{precision}}, \bar{\text{recall}}, \bar{\text{F2}}$
8 **Function** `DetermineOptimalThresholdOnTrainingSet():`
9 **foreach** *subject* ν *in* T **do**
10 **for** *threshold* $t = 0, ..., 1$ *in steps of* 0.01 **do**
11 $c_{i\nu} \leftarrow$ `PostProcPreds(`ν, t`)` ; `// Post-processing`
12 $\text{F2}_{\nu t} \leftarrow$ `Eval(CompareTruthPredPerSubject(`$c_{i\nu}$`))` ; `// Compare intervals`
13 $t_{\text{opt}} \leftarrow$ Get threshold t with the highest average $\text{F2}_{\nu t}$ over ν ; `// Find optimal F2`
14 **return** t_{opt} ; `// Return optimal threshold`
15 **Function** `PostProcPreds(`ν, t`):`
16 **if** *window-based classification* **then**
17 Convert $p_{\eta\nu}(\mathbf{x})$ to binary predictions $c_{\eta\nu}$ if $t \neq None$; `// Thresholding`
18 Resample $c_{\eta\nu}$ to get binary predictions $c_{i\nu}$ per data point ; `// Resampling`
19 **else** `//` Continuous segmentation
20 Convert $p_{i\nu}(\mathbf{x})$ to binary predictions $c_{i\nu}$ using threshold t ; `// Thresholding`
21 Merge intervals in $c_{i\nu}$ closer than δ ; `// Interval merging`
22 **return** $c_{i\nu}$; `// Return post-processed predictions`
23 **Function** `CompareTruthPredPerSubject(`c_i`):`
24 Init TP, FP, FN to zero and empty set M^{P} for tracking matched predicted intervals P in c_i;
25 Extend each true interval G by b^{before} and b^{after} ; `// Buffer ground-truth`
26 **foreach** *extended ground-truth interval* G **do**
27 **if** *at least one overlap of* G *with any* $P \notin M^P$ *exists with length(*P*)*$\leq d$ **then** `// Selecting`
28 Add first overlapping P to M^{P}; `// Track matched interval`
29 TP++ ; `// Increment TP`
30 **else**
31 FN++ ; `// Increment FN`
32 Set FP to the number of predicted intervals P not in M^{P} ; `// Count FP`
33 **return** TP, FP, FN ; `// Return TP, FP, FN`

Selecting appropriate metrics is crucial for evaluating and analyzing model performance. Following the taxonomy by Sørbø and Ruocco (2023), we select the F2 score as the final metric for optimization (performance evaluation) and use precision and recall as auxiliary metrics for additional insights (performance analysis). ALPEC computes the micro-average F2 scores across all subjects in the training set T to determine the optimal decision threshold t_{opt} (line 13). This threshold is then used to calculate the metrics

for each subject in the validation set V, which may also be the test set (line 6). Results are aggregated using the mean to adequately represent individual outliers (line 7).

Appendix G. Overcoming evaluation pitfalls with ALPEC

Authors employing window-based evaluation often overlook reporting the class balance between arousal and non-arousal samples. In instances where the balance is disclosed, such as in the work of Kuo et al. (2023), who reported a ratio of 42,311:33,479 (arousals vs non-arousals), and Badiei et al. (2023), whose confusion matrices implied a ratio of about 1:2, discrepancies arise. Our CPS dataset indicates an expected ratio of about 1:5 for 30-second windows, based on the average total sleep time and the number of arousals across subjects. Such disparities are problematic for comparative analyses and from a production standpoint, as they likely lead to underestimations of false positives when background samples are underrepresented. Our ALPEC framework addresses this by sampling at the subject level rather than the window level, ensuring that validation samples are representative of the overall dataset.

Moreover, the lack of cross-subject validation is a frequent oversight with window-based evaluations, where samples (intervals) from all subjects are often mixed across training, validation, and test sets. Since production models are applied to unseen subjects, it is critical to evaluate these models on new subjects during development. This practice is not consistently reported, which can inflate perceived model performance. ALPEC inherently avoids this issue by enforcing subject-level sampling, ensuring that the division of training, validation, and test samples maintains subject integrity. This approach enhances the comparability of results across studies and provides a more authentic evaluation of model efficacy.

Appendix H. CPS dataset details

A detailed description of all channels and fields within the dataset, translations of data fields from German to English, and Croissant (Akhtar et al., 2024) metadata (under the Apache-2.0 license) are provided on the PhysioNet page of the CPS dataset (Kraft et al., 2024). To avoid redundancy and potential ambiguities in case of updates on the PhysioNet page, we have not included this information in this paper. All relevant information about the data used in in the main part of this paper is however described in the following.

Inclusion and exclusion criteria Patients included in the dataset were aged 18 or older and referred for polysomnographic examination at a sleep laboratory. Patients undergoing diagnostic treatments in the form of positive airway pressure therapies were excluded.

Data extraction and preprocessing The data extraction involved multiple steps using the SOMNOscreen device from SOMNOmedics GmbH, capturing a broad range of physiological signals. The data was further processed using the DOMINO software from the same manufacturer, which calculated additional data channels and provided initial annotations for sleep stages and arousals, which were manually reviewed and adjusted by medical experts from NRI Medizintechnik GmbH, according to guidelines from the American Academy of Sleep Medicine (AASM) (Berry et al., 2012). The raw data channels were upsampled to a uniform sampling rate of 256 Hz.

All input features used in this work are described in Table 15 (raw measurement data), Table 16 (derived channels), and Table 17 (nominal event data).

Additional preprocessing of the data channels before release involved shifting the day, month, and year of all recordings to January 1, 1970, to ensure patient anonymity and converting from the European Data Format (EDF) to the Waveform Database (WFDB) format.

The target arousal classes are listed in Table 17. The presence of the postfix *(EEG)* at an arousal event class indicates that the arousal was first recognized in the EEG channel, followed by its causative occurrence. In contrast, the lack of *(EEG)* denotes that the causative event preceded the observable EEG effects. Another class of arousals that is also annotated but not included in this work are autonomic arousals. This exclusion is based on the distinct nature of autonomic arousals, which involve involuntary physiological responses regulated by the autonomic nervous system, differing from arousals typically detected in sleep studies through EEG or related to specific sleep disturbances. Autonomic arousals may not directly correlate with sleep architecture changes or the specific arousal events typically analyzed in sleep medicine, thus requiring separate consideration from

Table 15: Raw data channels

Channels	**Description**
C4:A1, C3:A2, F4:A1, O2:A1, A1, A2, C3, C4, F4, O2	Electroencephalogram. Single electrodes mean that this electrode is derived against all other electrodes.
Battery	Battery voltage level
Motion	Movement sensor measuring patient's physical activity or motion
Pressure Flow	Airflow pressure measured using oxygen nasal cannula at the nose and mouth
Thermal Flow	Thermal airflow sensor measuring breathing flow rate
ECG 2	Electrocardiogram measuring heart's electrical activity
EMG+, EMG-, EMG	Electromyogram measuring skeletal muscle activity at the left side (-) and right side (+) of the chin
EOGl, EOGL:A1, EOGL:A2, EOGr, EOGr:A1, EOGr:A2	Electrooculogram measuring the left (l) and right (r) eye movements
Light	Ambient light sensor measuring light exposure
PLMl, PLMr	Periodic Limb Movement sensors measuring limb movements at the left leg (l) and right leg (r)
Pleth	Plethysmography measuring changes in blood volume at the tip of the ring finger of the non-dominant arm
Pos.	Body position sensor. Used to derive the patient's posture
Pulse	Pulse rate of the pulse wave
RIP.Abdom, RIP.Thrx, Sum RIPs	Respiratory Inductance Plethysmography sensors measuring abdominal and thoracic movements during breathing. *Sum RIPs* is a combination of *RIP.Abdomen* and *RIP.Thrx*
SPO2	Pulse oximetry sensor measuring blood oxygen saturation levels
Snoring Sound	Snore sensor measuring snoring sounds or vibrations
Snoring Pressure	Pressure sensor measuring snoring intensity using oxygen nasal cannula at the nose and mouth

Table 16: Derived signals which are calculated by the DOMINO Software from the raw data

Signal name	**Description**
Syst	Systolic blood pressure curve
Diastol	Diastolic blood pressure curve
MAP	Mean arterial pressure
Diastol PTT	Diastolic pulse transit time
Systol PTT	Systolic pulse transit time
SpO2	Average oxygen saturation level
Integrated EMG	Integrated electromyography signal from the chin
PTT Raw	Pulse transit time
HRV LF	Low frequency component of heart rate variability
HRV HF	High frequency component of heart rate variability
Heart rate	Heart rate curve
RR Interval	RR interval for heart rate analysis
SVB	Sympathovagal balance of sympathetic and parasympathetic activity
RR	Respiratory rate per minute
Obstruction	Obstruction curve in synchronized effort from abdomen and thorax
Phase Angle	Phase angle of synchronized effort
Alpha+Beta FFT	Alpha and beta wave frequency analysis in sleep
Delta FFT	Delta wave frequency analysis in sleep
Sigma FFT	Sigma wave frequency analysis in sleep
Average Frequency Value	Average frequency value in sleep FFT analysis
Activity	Activity level
Light	Light intensity in lux

Table 17: Annotated events that are used in this work. For a complete list of all annotated events, refer to the official CPS dataset documentation (Kraft et al., 2024).

Event name	**Description**
Respiratory Arousal (EEG)	EEG arousal due to respiratory effort
Respiratory Arousal	Arousal due to respiratory effort
Flow Limitation Arousal (EEG)	EEG arousal due to flow limitations
Flow Limitation Arousal	Arousal due to flow limitations
SpO2 Arousal (EEG)	EEG arousal due to oxygen desaturation
LM Arousal (EEG)	EEG arousal due to limb movements
LM Arousal	Arousal due to limb movements
PLM Arousal (EEG)	EEG arousal due to periodic limb movements
PLM Arousal	Arousal due to periodic limb movements
Snoring Arousal (EEG)	EEG arousal due to snoring
Snoring Arousal	Arousal due to snoring
Arousal (EEG)	Spontaneous EEG arousal
Arousal	Spontaneous arousal
Sleep Profile: N1	N1 sleep stage
Sleep Profile: N2	N2 sleep stage
Sleep Profile: N3	N3 sleep stage
Sleep Profile: Rem	Rapid Eye Movement sleep stage
Sleep Profile: Wach	Awake state during the measurement
Body Position: Prone	Prone body position
Body Position: Upright	Upright body position
Body Position: Left	Lying on the left side
Body Position: Right	Lying on the right side
Body Position: Supine	Supine body position
Hypopnea	Hypopnea event
Apnea	Apnea event
Central Apnea	Central apnea event
Obstructive Apnea	Obstructive apnea event

a sleep medical perspective. Autonomic arousals are also typically not included in other ML-based works which focus on the general arousal detection task.

Loading the dataset The offical documentation of the CPS dataset on PhysioNet (Kraft et al., 2024) contains code files and instructions on how to load the dataset based on Croissant specifications (Akhtar et al., 2024). These are also attached in the supplementary material. Use instructions are provided in the *README.md* file.

Dataset statistics and analysis of representativeness Table 18 provides an overview of the CPS dataset, including demographic information, sleep architecture, and sleep disorder indices.

Since the CPS dataset was collected over the course of one year during routine clinical practice, it is expected to be representative of patients undergoing polysomnographic examinations in a sleep laboratory. Our analysis reveals that close to 80% of the patients in our study suffer from obstructive sleep apnea (OSA) of varying severity. The dataset features a male-to-female ratio of 45:17 (noting that gender information is not available for all patients), with about 70% of the patients being over 50 years old, and over 40% classified as obese (BMI $>$ 30). These characteristics align with findings in existing literature, which indicate that OSA is more prevalent among males, older individuals, and those with obesity. Such demographic patterns are well-documented in research, supporting the representative nature of the CPS dataset (Bonsignore et al., 2019; Jehan et al., 2017; Fietze et al., 2019).

The official CPS dataset page on PhysioNet and the supplementary material include a script named *generate_statistics.py* that uses the data loading functions described in Section H to load the data. This script generates basic demographic statistics, statistics on questionnaire answers, medical diagnoses, and additional derived statistics (e.g. distribution of the number of arousals per subject). All statistics are automatically generated using the SweetViz (Bertrand, 2020) library, available under the MIT license. Precomputed statistics are provided in the file *statistics.html*.

Appendix I. Data folds

The split between training and validation samples is performed randomly. The test set, however, was hand-selected by an expert in arousal diagnostics whose task was to find a set of samples that is representative of patients undergoing polysomnographic examinations in a sleep laboratory. A mapping of sample IDs to folds can be found in Tables 19 to 21. A few samples are excluded in our experiments: Five entries, marked with †in the training and validation folds, since they do not contain *Diastol* and *Syst* derived channels and one entry, marked with †in the test set, since it contains overlapping target annotations.

Appendix J. Datasheet

The following datasheet is based on the *Datasheets for Datasets* framework Gebru et al. (2021).

J.1. Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
The CPS dataset was compiled to conduct a clinical study on arousal diagnostics (Wienhausen-Wilke and Kraft, 2024). The primary study goals are to investigate if Machine Learning can enhance the quality and efficiency of sleep-related arousal diagnostics, while also reducing technical demands. The dataset was created with the specific task of refining the diagnostic workflow by leveraging Photoplethysmography (PPG) data to reduce reliance on comprehensive EEG, electromyography (EMG), and electrooculography (EOG) inputs.

Who created the dataset (e.g., team, research group) and on behalf of which entity (e.g., company, institution, organization)?
The dataset was compiled from patients undergoing regular polysomnographic examinations at the sleep laboratory of the Klinik für Kardiologie, Pneumologie und Angiologie at Klinikum Esslingen. The companies IT-Designers Gruppe and NRI Medizintechnik GmbH were involved in collecting and processing the dataset. IT-Designers Gruppe initiated, funded, and supported the research. NRI Medizintechnik GmbH operates the sleep laboratory and cooperated and assisted in implementing the data collection protocol. Technical support was provided by SOMNOmedics GmbH, the supplier for the hardware and software of the sleep laboratory. The clinic and companies are all based in Germany.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grant and the grant ID.

Table 18: **Characteristics of the CPS dataset.** The total number of patients is 113. The number of patients who indicated their gender in the questionnaire was 62, no other genders were mentioned. The number of patients with REM sleep was 110.
Data are expressed as absolute values or as mean values ± standard deviation in the units provided.
OSA Severity is determined based on the AHI (Wetter et al., 2012), where *Very Severe* is an additional category reserved for patients with *Severe* OSA and additionally a high hypoxemia burden and high daytime sleepiness with a tendency to fall asleep during the day. The Baveno Classification for OSA severity (Randerath et al., 2018) was newly introduced during the collection of the CPS dataset.
Abbreviations: BMI: Body Mass Index, TST: Total sleep time; WASO: Wake time after sleep onset; Sleep Stages: N1, N2, N3: Stages of non-rapid eye movement sleep, REM: Rapid eye movement sleep; OSA: Obstructive sleep apnea; AHI: Apnea–hypopnea index; ArI: Arousal index; ESS: Epworth Sleepiness Scale; SPO_2: Oxygen saturation level; ODI: Oxygen desaturation index ($\geq 3\%$); T90: Time percentage spent below 90% oxygen saturation during sleep.

Age (years)		**OSA Severity**	
<50	26 (23.01%)	Mild	10 (8.85%)
50-60	34 (30.09%)	Moderate	39 (34.51%)
60-70	23 (20.35%)	Severe	22 (19.47%)
>70	22 (19.45%)	Very Severe	17 (15.04%)
Unknown	8 (7.08%)	Other	25 (22.12%)
BMI (kg/m²)		**Baveno Classification**	
18.5-25	19 (16.81%)	Type A	15 (13.27%)
25-30	43 (38.05%)	Type B	32 (28.32%)
>30	48 (42.48%)	Type C	8 (7.08%)
Unknown	3 (2.65%)	Type D	1 (0.88%)
Gender		Unknown	57 (50.44%)
Male	45 (72.58%)	**Mean ESS**	7.83 ± 4.96
Female	17 (27.42%)	**Mean Number of Arousals**	167.56 ± 87.17
Sleep Architecture		**Sleep Disorder Index**	
Sleep Efficiency (%)	70.04 ± 15.87	AHI (events/hour)	107.86 ± 39.60
TST (min)	435.48 ± 36.83	ArI (events/hour)	20.95 ± 10.42
WASO (min)	130.45 ± 69.35	Snoring Index (events/hour)	65.25 ± 106.53
N1 (% of TST)	15.74 ± 11.70	**Oximetry Parameters**	
N2 (% of TST)	51.22 ± 11.58	SPO_2 (%)	93.75 ± 1.67
N3 (% of TST)	19.31 ± 11.62	ODI (events / hour)	22.23 ± 15.63
REM (% of TST)	13.74 ± 6.40	T90 (%)	8.04 ± 11.66

Table 19: Training set


```
3DquDEk2YwjfckxNBAQuVTshrK3VWqO7
HvVu33fnVKDLLjwY8Mtytcgi8Btsr1kS
INzmELsQB5yeF6HnHRM76U1ufVy7vmfb
hOwipKAoUqK6vJDqsjnchMKZf0e9uSH8
5Nl24yHOnojhw7nsgCOe530b2RBzOuLA
1RLVkOocGDZLI8RRhPglAc4I3gMSLqvu
0Ah95Qw18puf1JsnrKBA6u8XXZLlMIQJ†
Bmy6KwUhfqRDp6bzRx1PaWoQvBpImF01
FddiLTFWMZFHH5s1NFddllezef4BJhwS
tfAnzkFia5hzaA6bHYFpkj3jPF90FzAj
tU3dZpxIdmbr9wpPpFeZGh5MciOB1TgT
FPSnBoS217CEJ8cZS7MO3VuYUJwIt8LV
5wPINWASVdhb63RK4DtJt5LuyuaWyMo8
JyXyuQIuFyAtKL8ZoWS98xvpW4PJFco6
KIdPVCRCkXIaWDQ4c4gU38xpH5PAnOSV
CMTsLOEWEJvQqMe0GKKCKN87IXp9LOUU
0pdaTB9613iRUlhUJRAYvKmMQcOV3TYn
1gQ3otWoJ3qNNJ5g4N1WtTC4JTFOlP9B
wl5eUqFCFGibvn13l8axb82mNOkp0doc
DlUvX2vg1c4ORh63BRFaCxbMlr8DeCbO
8OLORVUBBBcTBzaPxmmnfrlXC3bu3Dzw
rxODabL5H6LkFhXhov0iCYKaxTA9SKSY
MYzsHdeN4rEc6ne1SobyUoK2u2bedJUp
sUDb7jmMMO6h7QCGGbkwa7LRv1JYu8hy
5C1M33g2KLtBshvnj3V6S2MYFxbvFgbr
S8dOGQgMx9WOH90orySWSfGuBL2mpxgC
UldfKBDGVlNpqUbcVPLV68eOFqucUjO6
zjSqovtw62tjDA2tSelpFpAvfXbGUyI5
dxAGfKahLvCc0GhMyac32CmQo10JGLuF
3rt8nhT9Ddda15KAPVhRXJgPcilmGipU
5mhIirR785Vve6LjBZyOqRviHmWUZl9M
kxrBCJhec2Aub2FmnrU1dCIx7f3HlST6
rYbtzQzJJVg8Wksx9dU8wHg2X1R8zulJ
yAP2PJs1dDSFdYXe6GrQtQC7iO8oyX3L
hrBQwXe1RNG3VJXIAQAgC7JAjl2XlHyD
PKOt8NcGbxRcfL8tTxbrdxJe6zY1WS9f
FRU8DMa3f1esZLFfzixxhgJkOKvG7vXX
8XdpUGiGQTqm1q6E1BiNZmYTcFwQovHP
1vxo6QPCl4cn7JrPoKgTFUd6zZ6jrN6X
HeeJJy8N63XTT1mmruOcaXwl9gHO6LLR
MyE1OSaDFTs03JE5K1CLusYomSaDFufz
l1VJ9A1olk859kfiGM7UNdjqitFHl3fa
eXOCkhjYbsXlnwQwAHsFDo9XyPV9b7oj
ZizO4wnchcm4tTOACaPUrkRtopGIjkFq
P45k7fhnHTEWYkEMfEPSAK2tp2hizA7O†
mkcvU3fDfRGULYMwqmOFCUlqPTzrSb8P
RhDBHQQCPFEvVYagp1SorSLbEygFUiAL
C2Lt7JOOpGRwdu2TrCMDNn1jE8nBCHpb
KkgHbcRejP3vgmwPVpI3jW3Pq6cddRsJ
F8Uu1bzONcullMqx3o7S1o3M4VRg7EtR
52JOaQM1ksTl9M7xo3AYi3Hto6exeWpG
nozzxBTVeanjyN8aDAOknM5gv55sydOG
t1dRq19eE7PizNhNBod8AT5pX18KFAul
Yp6NRNtFSjbjdcFWN7Q1ATwOir9wrBNw
Oa7cBsB7PQPy7Kr8jHkJorml61mv4kIP
srARDl4a3Z4Gtb39GTOv0ynNVE6T7xHS
KX83zwqAkUKsTWQKeWmfjewObf8y7upu
wuxfEVb6iJ7GghVUjN5eKTO0VsulXaOg
gm7PhGPwaWaEeOoG0lK0altS9tLOBYGt
zWtiCjFSxBFRmU3DklC4UMFKFHCOXJgS
xc54eJI7x5BwwtOLSfdRnYdvMLagr1sp
RqC1HNwuWs5fhqEheSYlRc93EoReggVm
U2onxpoiCTT2F4SHmDMTeWb2GgWtRZhb
Adm1jCIZpVR5Evov0BZ5XyEA3QXgkpiO
niVBznqEcE3RfyWa2u7EXjpguXBMB9dE
nrtZyE7IBmlm0ZH3gG7FHJAoYxlFqczB†
x6nbOt0fqHhNHbYCgQPOEaa7ymp6eKuq
QHauJpImWK54mPysmthbCSRI6BwQfuim
```


Table 20: Validation set


```
I4PBCtY88EMiljTpJ6ns5kIoimlIUgfl
ip25KNFw4RbDNeQTXjLI95OsnQLLe35e
vvcikOyQkXvfZdNkFYEWveZjUlQmYrSf
euzXQcpDnB1xsKeKOJHYW48lSNZXiVHK
AavMakHhFeoHz9AgdWVTBsXCdiOLBNEk
KBn6Fz4XRNh5A7YRBBGZIlSVzcmV6P51
AEhNZrB5mb9K7hCBxBOxvJzUC5WxqaFN
ssRHUDqjbEgtepnEnjPpOd6WQHekD6PZ
XpSmjNEWOMad8655KQ4q3NjDHNUNHW6C
9q7wGfr4xVuiodCCrqViuS1dnZ8tthvZ
A2YTNgCrkIoMGvSkyzK5ROEFgXVRfptP
xijlZSyZXb5MEfbKLM54iOJRJkRdrBrX
OCi1DrQQgJ7GjNFNTrxscclxkaDHOY76
yWLp2YwYQlKqoSR14JUiazZzcywX8Xj8
jusHjPORbNBFg6iMvaNE4HcFMc4oJyqD
3veUj2KxK6jmJlllRj9tXbSncK6xnSOx
ZwojnSyDEhD6s8ZERY8AoDWL0cnBH7BU
GJShIRZ3lsuqeG7HhgpwlMw5v8prIuD4
ZFkXoBHziu00Toye78g1YdKm5P0bA0dY
pY4EUZQszT0kL5Et76FkEbRlLeAMT2hz
oTlKy2ISbZfN7i2jTcZ4mCqmCpK6dhDm
TLYaLIWCrBbtqLySvqxj67g5ZAfn2Zhs†
MT9Dkn3L0akMVov0lRUq9HmNsP49R1dX
rHg2gQNoevGYPUag6PAn9CANXKmx17ms
Vth5VKxswvoQEpLCis200xjS0f0ctjFu
vc3ShhPeF5CiTm9Hi0moakicyNWNulab
XSedsGunPd3IUuZ8RJUGmz3SUoCv0rzW
Pk9lFIEExot3gjvV83ZDTuWSj6dqk9uW
tihrah4T4i2gA8dscroj9Mu5715fXojg
M7d3C4ZQtxOR0OWum7JMxQtZExUfNEzi†
```


Table 21: Test set


```
vHJMSYFIl1TfLweQ5DWMGN5f47ULFNxe
MT1zW5iBOh1bxF42QBpyDqotQk7NcnHw
leySrSnra9yAO3eTJIGB55nrjRS3RqIW
FLsgQZoIGHx1G3LmdD7jtICMik2EKRKN
YFQX33c8EEoapTndd2084KbUuUmtj7xF
KB84bUmLWOrKCKkISCn8QuNBhF5mgOL8
oEJ7fslCTL7s0OfIe7nYIPqo7Il4rMjI
CzwqE37s81YahjNICSXI2Tb4Fmp6bclE
Su02hndUSYGSKJmcSqroKmtDjXIJ4y60
LcsapTberZwzU7qyEr11andO59HTOVCv
0vjSLbBj2sckqQam3tZ92QLDpQNYqaa0†
mXHZZ887A9fcZgOmnxhnPVHwu5ECljDG
RpARZ17l5osnFUcqIj2aTOsgRBMDutoA
tIgyhF8T1BOZnu7h6jb58igU5MAGgdo9
kKDzUlAprXqDz84Nrw9UP1WOjpgUKkhN
```

This research was funded by STZ Softwaretechnik GmbH, part of the IT-Designers Gruppe, Esslingen am Neckar, Germany. It contains all samples that have been collected during the clinical study.

Any other comments?
None.

J.2. Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., nodes, edges) present in the dataset? Please provide a description.
The instances in the dataset represent diagnostic polysomnographic sleep recordings, which include up to 36 raw and 23 derived data channels, alongside 81 types of annotated events for each participant, supplemented by data from various questionnaires.

How many instances are there in total (of each type, if appropriate)?
The dataset encompasses 113 diagnostic polysomnographic sleep recordings.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not.
The dataset is a sample of diagnostic sleep recordings from adult patients undergoing regular and purely diagnostic examinations at the sleep laboratory of the Klinik für Kardiologie, Angiologie und Pneumologie of the medical clinic in Esslingen am Neckar, Germany. It contains all samples that have been collected during the clinical study. The representativeness compared to other datasets on arousal diagnostics is discussed in Appendix H.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
Each instance consists of raw polysomnographic data including up to 36 raw and 23 derived data channels, 81 types of annotated events, and data from various questionnaires. Descriptions of all channels and fields that are used in this work are provided in Appendix H. A full description of the dataset is available on the PhysioNet page of the CPS dataset (Kraft et al., 2024).

Is there a label or target associated with each instance? If so, please provide a description.
Yes, each recording includes labels for sleep-related events such as arousals, apnea, hypopnea, and other sleep events as per the American Association of Sleep Medicine (AASM) guidelines (Berry et al., 2012). A full list of annotated events is available on the PhysioNet page of the CPS dataset (Kraft et al., 2024).

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
The Pittsburgh Sleep Quality Index (PSQI) questionnaire was only given to 62 patients. The remaining 51 patients did not receive the questionnaire, so this information is missing for those instances. All questionnaires contain missing values for some questions due to non-response. Raw data is complete for all patients, but five patients are missing derived systolic and diastolic blood pressure channels.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.
No explicit relationships between individual instances are made.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
We provide recommended splits for training, validation, and test data, listed in Appendix I and on the PhysioNet page of the CPS dataset (Kraft et al., 2024).

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
The dataset may contain some noise typical of polysomnographic data, such as artifacts from patient movement or external interference, but efforts were made to minimize these by conducting comprehensive quality assurance in the pilot phase of the clinical study. The channels *Heart Rate*, *Light*, and *SpO2* are smoothed versions of the raw data channels *Pulse*, *Light*, and *SPO2*, respectively, processed using the DOMINO software from SOMNOmedics GmbH.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If external resources are required, please describe them, as well as any restric-

tions (e.g., licenses, fees) associated with them.
The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
Yes, the dataset contains confidential patient data protected under doctor-patient confidentiality agreements. The most sensitive data are the sleep medical diagnoses. The dataset has been anonymized to protect patient identities.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
No, the dataset does not contain such data.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
The dataset includes demographic information such as age and gender to study their impact on sleep-related arousals. It was determined from questionnaires. The age distribution in years is: Under 50: 26 patients (23.01%), 50-60: 34 patients (30.09%), 60-70: 23 patients (20.35%), over 70: 22 patients (19.45%), unknown: 8 patients (7.08%). The approximate gender distribution, based on 62 patients, is: Male: 45 patients (72.58%), Female: 17 patients (27.42%). Gender information is only available in the aggregated statistics, not for individual patients, as an anonymization measure.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
No, patient identities are anonymized to prevent identification.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, locations of health data about individuals or genetic data, forms of financial information, such as social security numbers, salary)? If so, please provide a description.
The most sensitive data are the sleep medical diagnoses, consisting of short extracted textual descriptions of patients' sleep-related medical conditions from doctors' letters, e.g., obstructive sleep apnea with severity indications. Statistics and a complete listing of the sleep medical diagnoses are available in the official documentation on the PhysioNet page of the CPS dataset (Kraft et al., 2024) and in the file *statistics.html* in the supplementary material (generated using the SweetViz (Bertrand, 2020) library, under MIT license).

Any other comments?
None.

J.3. Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie rating), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part of speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
The data was directly observable from polysomnographic recordings and supplemented by data from various questionnaires.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?
The data collection during polysomnographic examinations involved SOMNOscreen devices and the DOMINO software, both from SOMNOmedics GmbH, Germany. The data was validated, curated, and extended with additional labels (e.g., sleep stages and arousals) by trained sleep medical scorers from NRI Medizintechnik GmbH (Germany) following guidelines from the American Academy of Sleep Medicine (AASM). The data was exported from DOMINO in EDF (raw data) and TXT (annotations) formats. An employee of Klinikum Esslingen (funded by IT-Designers Gruppe) performed the digitalization of the questionnaires and doctor's letter in YAML-format, the pseudonymization of the whole data and made the data available to the research group from IT-Designers Gruppe. The data was then further anonymized for release. The whole process was developed in collaboration with data protection officers from Klinikum Esslingen and IT-Designers Gruppe. It was approved by the ethics committee of the Landesärztekammer Baden-Württemberg,

Germany. The data quality was validated in a pilot phase.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
The dataset was collected monocentrically at the sleep laboratory of the Klinik für Kardiologie, Angiologie und Pneumologie in Esslingen am Neckar, Germany, ensuring a representative sample of adult patients undergoing regular diagnostic examinations.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
The data collection was conducted by clinical staff at Klinikum Esslingen. One student was hired by the clinik to perform the digitalization and pseudonymization of the data. He was paid 12.98 EUR per hour. The expenses for the student were covered by STZ Softwaretechnik GmbH.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances? If not, please describe the timeframe in which the data associated with the instances was created.
The data was collected during 2021-2022.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
The study protocol was approved by the ethics committee of the Landesärztekammer Baden-Württemberg, Germany, on 2020-10-21 (committee number F-2020-105, https://www.aerztekammer-bw.de/ethikkommission).
The clinical study was registered at the German Clinical Trials Register, DRKS-ID: DRKS00033641 (Wienhausen-Wilke and Kraft, 2024).

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
Yes, the dataset relates to people.

Did you collect the data from individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
The data was collected directly from individuals at the Klinikum Esslingen sleep laboratory.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
Patients were informed as part of the clinical study consent process. The consent form followed a template and was approved by the ethics committee of the federal state, the Landesärztekammer Baden-Württemberg, Germany.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
Yes, patients gave informed consent for the collection and use of their data. The content of the clinical study and all general conditions were explained verbally by the medical staff and in writing in the informed consent form. It was administered in a preliminary visit prior to the examination.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to, or otherwise reproduce, any supporting documentation.
Patients were informed of their right to revoke consent at any time.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
Yes, an impact analysis was conducted with a focus on the risks of re-identification, data misuse, data breaches, and failure to achieve the study objectives. Outcomes (measures) to reduce the risks were as follows: Selection of an established platform for data sharing (PhysioNet) including usage limitations (future usage must be in line with the original study goals) and the requirement of a data use agreement and public credentialed access. Additionally, in order to anonymize the data, we removed most free text, the sensitive attributes medication and preexisting conditions, and multiple indirect identifiers like gender and profession that were less important for achieving the study goals. For the remaining indirect identifiers (age and BMI), we selected bins that respected k-anonymity with k=3. For the medical diagnosis (the remaining sensitive attribute), we made sure to have i-diversity with i=2 among the k-

anonymous groups. Absolute timestamps in the measurement data were adjusted to start on January 1st, 1970. Our measures cover the criteria required by the safe harbor method from the Health Insurance Insurance Portability and Accountability Act (HIPAA) and go beyond them.

Any other comments?
None.

J.4. Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.
Derived data channels and some event data were automatically calculated by the DOMINO software from SOMNOmedics GmbH. Manual labeling was performed by trained sleep medical scorers from NRI Medizintechnik GmbH following guidelines from the American Academy of Sleep Medicine (AASM). Bucketing of age and BMI attributes and shifting of timestamps were performed to anonymize the data. Apart from this, all raw data were converted from EDF to WFDB format, which is the standard format for PhysioNet. In the process, all raw data was upsampled to the highest sampling rate of 256 Hz.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data? If so, please provide a link or other access point to the "raw" data.
Access to the raw data beyond the clinical study is limited to the medical clinic, Klinikum Esslingen.

Is the software used to preprocess/clean/label the data available? If so, please provide a link or other access point.
The software that was used for preprocessing is proprietary. The DOMINO software from SOMNOmedics GmbH can be licensed from the company.

Any other comments?
None.

J.5. Uses

Has the dataset been used for any tasks already? If so, please provide a description.
This publication entails the first use of the dataset.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
No.

What (other) tasks could the dataset be used for?
The dataset may be used for research that aligns with the original study goals (Wienhausen-Wilke and Kraft, 2024). The study is aimed at investigating the utility of Machine Learning (ML) for improving the quality and efficiency of sleep-related arousal diagnostics, reducing technical demands of the data collection process, assessing the utility of a transparent clinical decision support system, studying the clinical relevance of arousals on sleep quality, and the utility of ML for medical knowledge discovery in this context.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might limit its usability for other tasks?
We do not foresee any limitations on the usability of the dataset for the tasks mentioned above.

Any other comments?
None.

J.6. Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
Yes, the dataset is publicly available for research purposes with credentialed access on PhysioNet (Kraft et al., 2024).

How will the dataset be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
The dataset is available on PhysioNet (Kraft et al., 2024).

When will the dataset be distributed?
The dataset is already available on PhysioNet (Kraft et al., 2024).

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
The dataset is distributed under the PhysioNet Credentialed Health Data License 1.5.0[2].

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a

2. https://physionet.org/about/licenses/physionet-credentialed-health-data-license-150/

link or other access point to, or otherwise reproduce, any supporting documentation.
There are no third-party IP-based restrictions.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
There are no export controls or other regulatory restrictions.

Any other comments?
None.

J.7. Maintenance

Who will be supporting/hosting/maintaining the dataset?
The dataset is hosted on the PhysioNet platform (Kraft et al., 2024). Support and maintenance will be provided by the authors of this publication.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
Contact information for the dataset maintainers can be found in the official documentation on the PhysioNet page of the CPS dataset (Kraft et al., 2024) under the *Corresponding Author* section.

Is there an erratum? If so, please provide a link or other access point.
Currently, there is no erratum. If the need for an erratum arises, the dataset can be updated on PhysioNet with semantic versioning.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?
Updates will be deployed as necessary to correct any errors. Communication will be done via PhysioNet.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time, or were they told that they could have their data deleted)? If so, please describe these limits and explain how these limits will be enforced.
Anonymized data will be retained indefinitely on PhysioNet. Access to pre-anonymized data is restricted to the IT-Designers Gruppe for four years after the end of the data collection, as communicated to study participants. Klinikum Esslingen will retain the original data in accordance with legal requirements. Deletion requests only affect pre-anonymized data. Since the patient IDs in the anonymization process were created randomly and not linked to any patient information, deletion of the anonymized data of a specific patient is technically not possible.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.
Older versions will be maintained on PhysioNet to ensure continuity and reproducibility of research.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description.
We are not aware of any mechanism on PhysioNet for users to contribute directly to the dataset. However, we are open to collaboration and will consider any requests for extensions or contributions.

Any other comments?
None.

Appendix K. Utilized compute and environmental impact

Experiments following the *DeepSleep* approach ran on single Nvidia GeForce RTX2080TI GPUs with 11 GB of memory. They are part of a workstation containing four GPUs. The workstation is equipped with an Intel Core i9-10900X CPU with 10 cores and 20 threads and 128 GB of DDR4 RAM. Experiments using models implemented in the *sktime* library (Löning et al., 2019) ran on the CPU of the same workstation.

Table 22 gives an overview of the average runtime for single experiments conducted for this publication, where we also indicate how often experiments were repeated to obtain confidence intervals.

In total, we have 391.17 GPU hours and 20.26 hours without GPU usage. From this, we conduct estimations of $kgCO_2eq$ for the GPU hours using the MachineLearning Impact calculator[3] presented in Lacoste et al. (2019). We use a factor of 0.4880 $kgCO_2eq$ per kWh for the electricity mix in Germany.

Total emissions are estimated to be 47.72 $kgCO_2eq$. Due to preliminary exploratory experiments and repetitions of experiments due to changes in requirements or fixes, we estimate the total emmission to be thrice as high, i.e. about 150 $kgCO_2eq$.

3. https://mlco2.github.io/impact#compute

Table 22: Runtime and count of experiments conducted for this publication.

	Experiment	Environment	Repetitions	Average runtime [h]
Main experiments on the CPS dataset	D4	GPU	5	10.8
	D3	GPU	5	6.5
	D2	GPU	5	0.7
	D1	GPU	5	0.7
	IndividualBOSS	CPU	5	0.18
	SupervisedTimeSeriesForest	CPU	5	1.4
	TimeSeriesForestClassifier	CPU	5	0.7
	SignatureClassifier	CPU	5	1.1
	SummaryClassifier	CPU	5	0.2
	Catch22Classifier	CPU	5	0.4
	RandomStratified	CPU	5	0.03
	RandomUniform	CPU	5	0.03
	Constant 1	CPU	1	0.03
	Constant 0	CPU	1	0.03
2018 PhysioNet Challenge	Target, IOD	GPU	5	11.6
	Target, POD	GPU	1	23,75
	Target, FED	GPU	5	6.8
	Non-target, IOD	GPU	5	10,6
	Non-target, POD	GPU	1	33,2
	Non-target, FED	GPU	5	11,9
MISC	Ablation study	GPU	30	0.65
	HP Tuning	GPU	44	0.38