MSED: A Multi-Modal Sleep Event Detection Model for Clinical Sleep Analysis

Alexander Neergaard Zahid, *Member, IEEE*, Poul Jennum, Emmanuel Mignot, and Helge B. D. Sorensen, *Senior Member, IEEE*

***Abstract*—Clinical sleep analysis require manual analysis of sleep patterns for correct diagnosis of sleep disorders. However, several studies have shown significant variability in manual scoring of clinically relevant discrete sleep events, such as arousals, leg movements, and sleep disordered breathing (apneas and hypopneas). We investigated whether an automatic method could be used for event detection and if a model trained on all events (joint model) performed better than corresponding event-specific models (single-event models). We trained a deep neural network event detection model on 1653 individual recordings and tested the optimized model on 1000 separate hold-out recordings. F1 scores for the optimized joint detection model were 0.70, 0.63, and 0.62 for arousals, leg movements, and sleep disordered breathing, respectively, compared to 0.65, 0.61, and 0.60 for the optimized single-event models. Index values computed from detected events correlated positively with manual annotations ($r^2 = 0.73$, $r^2 = 0.77$, $r^2 = 0.78$, respectively). We furthermore quantified model accuracy based on temporal difference metrics, which improved overall by using the joint model compared to single-event models. Our automatic model jointly detects arousals, leg movements and sleep disordered breathing events with high correlation with human annotations. Finally, we benchmark against previous state-of-the-art multi-event detection models and found an overall increase in F1 score with our proposed model despite a 97.5% reduction in model size.**

***Index Terms*—Computational sleep science, object detection, deep neural network.**

Manuscript received 19 July 2022; revised 6 December 2022 and 16 February 2023; accepted 22 February 2023. Date of publication 3 March 2023; date of current version 30 August 2023. This work was supported in part by Lundbeck Foundation under Grant R347-2020-2439 with supporting grants from Augustinus Fonden, Knud Højgaards Fond, Otto Mønsteds Fond, Reinholdt W. Jorck og Hustrus Fond, Stibo Fonden, and Vera & Carl Johan Michaelsens Legat. *(Corresponding author: Alexander Neergaard Zahid.)*

Alexander Neergaard Zahid is with the Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby 2800, Denmark (e-mail: aneol@dtu.dk).

Helge B. D. Sorensen is with the Department of Health Technology, Technical University of Denmark, Denmark.

Emmanuel Mignot is with the Stanford Center for Sleep Sciences and Medicine, Stanford University, USA.

Poul Jennum is with the Danish Center for Sleep Medicine, University Hospital of Copenhagen, Denmark.

Source code for training and inference is available at https://github.com/neergaard/msed.git.

Digital Object Identifier 10.1109/TBME.2023.3252368

I. Introduction

CLINICAL sleep analysis is currently evaluated manually by experts based on guidelines from the American Academy of Sleep Medicine (AASM) detailed in the AASM Scoring Manual [1]. The guidelines detail both technical and clinical best practices for setting up and recording polysomnographiess (PSGs), which are overnight recordings of various electrophysiological signals including electroencephalography (EEG), electrooculography (EOG), chin and leg electromyography (EMG), electrocardiography (ECG), respiratory inductance plethysmography from the thorax and abdomen, oronasal pressure, and blood oxygen levels. Based on these signals, expert technicians score and analyse the PSGs for sleep stages [wakefulness (W), rapid eye movement (REM) sleep, non-REM stage 1 (N1), non-REM stage 2 (N2), and non-REM stage 3 (N3)], and sleep micro-events summarized by key metrics, such as the number of apneas and hypopneas per hour of sleep (apnea-hypopnea indices, AHIs), the number of (periodic) leg movements per hour of sleep [(periodic) leg movement index, (P)LMI], and the number of arousals per hour of sleep (arousal indices, ArIs).

Arousals (Ars) are defined as abrupt shifts in EEG frequencies towards alpha, theta, and beta rhythms for at least 3 s with a preceding period of stable sleep of at least 10 s [2]. During REM sleep, where the background EEG shows similar rhythms, arousal scoring requires a concurrent increase in chin EMG lasting at least 1 s [1].

Leg movements (LMs) should be scored when there is an increase in amplitude of at least $8\,\mu$V in the leg EMG channels above baseline level with a duration between 0.5 s to 10 s [3]. A PLM series is then defined as a sequence of 4 LMs, where the time between LM onsets is between 5 min to 90 min [1], [4].

Apneas are generally scored when there is a complete ($\geq 90\,\%$ of pre-event baseline) cessation of breathing activity. The underlying cause can be either a physical obstruction (obstructive apnea) or due to an underlying disruption in the central nervous system control (central apnea) for at least 10 s [5]. When breathing is only partially reduced ($\geq 30\,\%$ of pre-event baseline) and the duration of the excursion is ≥ 10 s, the event is scored as a hypopnea if there is either a $\geq 4\,\%$ oxygen desaturation or a $\geq 3\,\%$ oxygen desaturation coupled with an arousal [1]. Sleep disordered breathing (SDB) here refers to the collective of apneas and hypopneas.

Several studies have shown significant variability in the scoring of both sleep stages [6], [7], [8], [9], [10], [11], [12] and

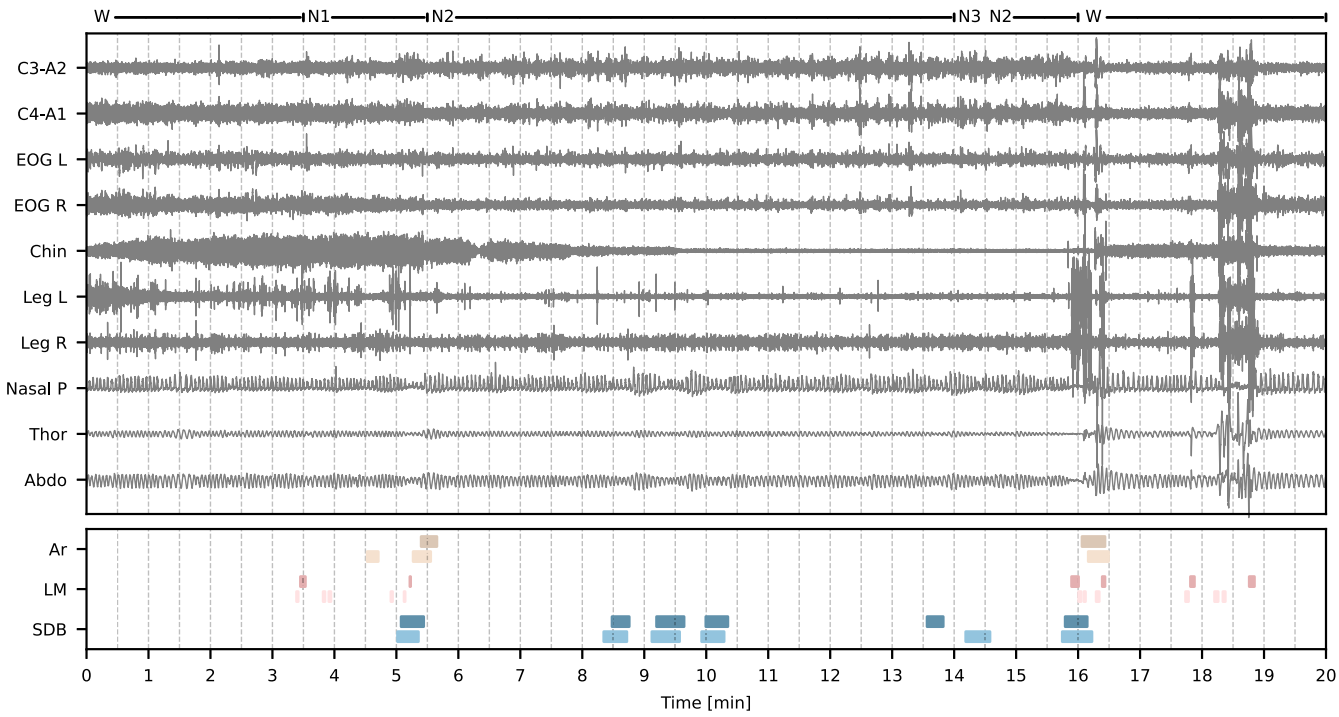

Fig. 1. Example of MSED input and output predictions. Ten PSG channels comprising EEG, EOG, EMG, and breathing are visualized, which are fed to the MSED model. The bottom shows the manually scored (darker shade) and model predicted (lighter shade) events associated with the given recording segment. Sleep stages are shown at the top. W: wakefulness; N1: non-REM stage 1; N2: non-REM stage 2; N3: non-REM stage 3; Ar: arousal; LM: leg movement; SDB: sleep disordered breathing.

sleep micro-events [5], [13], [14], [15], [16], [17], [18], [19]. This has prompted extensive research into automatic methods for classifying sleep stages in large-scale studies [20], [21], [22], [23], [24], [25], [26], [27], while the research in automatic arousal [28], [29], [30] and LM [31] detection on a similar scale is limited, but has attracted increased focus as evidenced by the *You Snooze You Win* PhysioNet challenge from 2018 [32], [33]. Biswal et al. recently proposed a multi-task CNN/RNN combination model for the purpose of classifying sleep stages and predicting apnea-hypopnea indices (AHIs) and leg movement index (LMI) [24]. They trained their model on 9000 PSG recordings from the Massachusetts General Hospital (MGH) and evaluated their model on a held-out MGH dataset consisting of 1000 PSGs, and on 5804 PSGs studies from the Sleep Heart Health Study (SHHS), yielding strong AHI/expert correlation values (0.85 on MGH, 0.77 on SHHS) and LMI/expert correlation (0.79 on MGH). Brink-Kjaer et al. published a CNN/RNN model for combined arousal and sleep/wake detection yielding an arousal detection F1 score of 0.79 on a test set of 1024 unique subjects [30], which was subsequently validated in two separate patient groups [34], [35]. Similarly, Carvelli et al. proposed a CNN/RNN model for LM detection reporting an impressive F1 score of 0.77 on 348 PSGs from the MrOS sleep study [31]. However, these models are all based on classifying windows of sleep data with subsequent manual fine-tuning and post-processing to combine events predicted in close-proximity windows, which incurs a human-factor bias.

Recently, Chambon et al. proposed the Dreem One Shot Event Detector (DOSED) algorithm for detecting sleep spindles and K-complexes in the sleep EEG [36], which was further developed for arousal and leg movement detection in subsequent publications [28], [37]. The advantage of this kind of approach is two-fold: first, the model is trained end-to-end to detect and classify events of any type, since there is no reliance on manual post-hoc processing of event predictions; and second, using a grid of default event windows (discussed in Section III-B) allows durations of different time scales. However, as these models were only designed for either EEG-only events [36], [37], or did not investigate joint detection of events occurring across multiple modalities [28], there is still an unmet need for models capable of predicting events of multiple classes from multiple sensor types.

In this study, we extend the previous work in [28], [36], [37] and introduce the multi-modal sleep event detection (MSED) model for joint detection of sleep micro-events. The model combines multiple recording modalities from the PSG and recent advances in machine learning to not only classify arousals, LMs, and SDBs, but also annotate them in the temporal domain without the need for post-hoc processing of predictions. An example of model predictions for a segment of PSG data is shown in Fig. 1, where the input signals are shown in the top box and manually scored/predicted events in the bottom box.

Our contributions are as follows:

- We propose the MSED, which is a CNN/RNN based model using disentangled feature extraction streams trained end-to-end for for multi-modal sleep event detection. To our knowledge, this is the first time this has been shown for multiple event types with multiple modalities trained simultaneously.

TABLE I
MROS DEMOGRAPHICS BY SUBSET

	$\mathcal{D}_{\text{train}}$	$\mathcal{D}_{\text{eval}}$	$\mathcal{D}_{\text{test}}$	p-value
n	1653	200	1000	-
Age, years	76.4 ± 5.6 [67.0, 90.0]	76.8 ± 5.4 [68.0, 90.0]	76.4 ± 5.3 [67.0, 90.0]	0.404
BMI, $\mathrm{kg\,m^{-2}}$	27.3 ± 3.9 [16.0, 47.0]	27.0 ± 3.6 [19.0, 40.0]	27.0 ± 3.7 [17.0, 45.0]	0.247
TST, min	357.3 ± 69.0 [54.0, 615.0]	354.0 ± 69.1 [108.0, 503.0]	353.6 ± 68.7 [62.0, 572.0]	0.312
SL, min	22.9 ± 25.6 [1.0, 349.0]	21.6 ± 23.0 [1.0, 135.0]	25.1 ± 32.1 [1.0, 402.0]	0.284
REML, min	109.5 ± 77.9 [0.0, 578.0]	103.5 ± 70.0 [10.0, 413.0]	107.2 ± 75.3 [3.0, 590.0]	0.466
WASO, min	116.7 ± 67.1 [11.0, 462.0]	119.0 ± 70.8 [15.0, 372.0]	112.9 ± 65.0 [6.0, 458.0]	0.471
SE, %	75.9 ± 12.1 [17.0, 97.0]	75.5 ± 12.3 [37.0, 96.0]	76.4 ± 11.8 [26.0, 98.0]	0.690
N1, %	6.8 ± 4.1 [0.0, 31.0]	7.0 ± 4.5 [0.0, 28.0]	6.9 ± 4.7 [1.0, 58.0]	0.968
N2, %	62.7 ± 9.5 [28.0, 89.0]	62.0 ± 9.7 [30.0, 90.0]	62.8 ± 10.0 [21.0, 95.0]	0.451
N3, %	11.4 ± 9.0 [0.0, 55.0]	11.8 ± 9.7 [0.0, 55.0]	11.1 ± 9.0 [0.0, 57.0]	0.638
REM, %	19.2 ± 6.5 [0.0, 44.0]	19.4 ± 7.2 [0.0, 41.0]	19.3 ± 6.7 [0.0, 42.0]	0.894
ArI, $\mathrm{h^{-1}}$	23.5 ± 11.8 [3.0, 87.0]	23.4 ± 11.0 [4.0, 77.0]	23.8 ± 11.8 [4.0, 102.0]	0.661
AHI, $\mathrm{h^{-1}}$	13.5 ± 13.9 [0.0, 83.0]	13.6 ± 13.3 [0.0, 59.0]	14.2 ± 15.5 [0.0, 89.0]	0.907
PLMI, $\mathrm{h^{-1}}$	35.4 ± 37.1 [0.0, 233.0]	36.6 ± 39.0 [0.0, 178.0]	36.0 ± 37.7 [0.0, 175.0]	0.993

BMI: body-mass index; TST: total sleep time; SL: sleep latency; REML: REM sleep latency; WASO: wake after sleep onset; SE: sleep efficiency; N1: non-REM stage 1; N2: non-REM stage 2; N3: non-REM stage 3; REM: rapid eye movement; ArI: arousal index; AHI: apnea-hypopnea index; PLMI: periodic leg movement index.

- We report inceased F1 scores using MSED compared to previous state-of-the-art in multi-event detection, despite a 97.5% reduction in memory footprint as defined by the number of model parameters.
- Clinically relevant endpoints as computed by MSED correlate strongly with expert-scored values.
- Source code for training and testing models are available at https://github.com/neergaard/msed.git.

II. DATA

We collected PSGs from the MrOS Sleep Study conducted between 2003 and 2005, an ancillary part of the larger Osteoporotic Fractures in Men Study [38], [39], [40]. The main outcome of the MrOS Sleep Study was to investigate and discover connections between sleep disorders, skeletal fractures, and cardiovascular disease and mortality in community-dwelling older (> 65 years). Of the original 5994 study participants, 3135 subjects were enrolled at one of six sites in the USA for a comprehensive sleep assessment, while 2909 of these underwent a first visit full-night in-home PSG recording. The PSG studies were subsequently scored by certified sleep technicians according to the prevailing guidelines at the time. Sleep stages were scored into stages 1, 2, 3, 4, wakefulness, and REM according to Rechtschaffen and Kales (R&K) scoring guidelines. For the purpose of this study, sleep stages were converted into their AASM equivalents (stage 1 into N1, stage 2 into N2, and stages 3 and 4 into N3) [1]. Arousals were scored as abrupt increases in EEG frequencies lasting at least 3 s according to older guidelines from the American Sleep Disorders Association [41]. Apneas were defined as complete or near complete cessation of airflow lasting more than 10 s with an associated 3 % or greater SaO_2 desaturation, while hypopneas were based on a clear reduction in breathing of more than 30 % deviation from baseline breathing lasting more than 10 s, and likewise assocated with a greater than 3 % SaO_2 desaturation. While the scoring criteria for scoring LMs are not explicitly available for the MrOS Sleep Study, the prevailing standard at the time of the study was to score LMs following an increase in leg EMG amplitude of more than 8 µV above resting baseline levels for at least 0.5 s, but shorter than 10 s [42]. A subset of the 2909 subjects also participated in follow-up sessions, although these studies did not include scoring of leg movements.

A. Subset Demographics and Partitioning

We used all first visit PSG studies ($n = 2853$) available from the National Sleep Research Resource (NSRR) [43], [44], which we partitioned into a training set ($\mathcal{D}_{\text{train}}$, $n_{\text{train}} = 1653$), a validation set ($\mathcal{D}_{\text{eval}}$, $n_{\text{eval}} = 200$), and a final testing set ($\mathcal{D}_{\text{test}}$, $n_{\text{test}} = 1000$). Key demographics and PSGs-related variables for each subset are shown as mean ± standard deviation with range in parenthesis in Table I.

1) Signal and Events: For this study, we considered three PSG events: Ars, LMs, and SDB events. These types of events are based on a specific set of electrophysiological channels from the PSG consisting of left and right central EEG (C3 and C4), left and right EOG, left and right chin EMG, left and right leg EMG, nasal pressure, and respiratory inductance plethysmography from the thorax and abdomen. EEG and EOG channels were referenced to the contralateral mastoid process, while a chin EMG was synthesized by subtracting the right chin EMG from the left chin EMG.

Apart from the raw signal data, we also extracted onset time relative to the study start time and duration times for each event type in each PSG.

III. Methods

A. Notation

We denote by $[\![a, b]\!]$ the set of integers $\{n \in \mathbb{N} \mid a \leq n \leq b\}$ with $[\![N]\!]$ being shorthand for $[\![1, N]\!]$, and by $n \in [\![N]\!]$ the nth sample in $[\![N]\!]$. A segment of PSG data is denoted by $\mathbf{x} \in \mathbb{R}^{C\times T}$, where C is the number of channels and T is the duration of the segment in samples. An event type is defined as $\varepsilon_i = (\varrho_i, \delta_i, l_i) \in \mathbb{R}_+^2 \times \mathcal{L}$, where ϱ, δ, l is center point, duration, and label of the event, and $\mathcal{L} = [\![L]\!]$ is the event label space. The set of N_t true events for a given PSG segment is denoted by $\boldsymbol{\varepsilon}^t = \{\varepsilon_i^t \mid i \in [\![N_t]\!]\}$. By $\chi \in \mathcal{D}_*$ we denote a sample in either one of the three subsets. In the description of the network architecture, we have omitted the batch dimension from all calculations for brevity.

B. Model Overview

Given an input set $\boldsymbol{\chi} = \{\mathbf{x}, \boldsymbol{\varepsilon}^t\} \in \mathbb{R}^{C\times T} \times \mathbb{R}_+^{N_t\times 2} \times \mathcal{L}$ containing PSG data with C channels and T time steps, and true events $\boldsymbol{\varepsilon}$, the goal of the model is to detect any possible events in the segment, where, in this context, detection covers both classification *and* localization of any event in the segment space.

The model generates a set of *default event windows* $\boldsymbol{\varepsilon}^d = \{\varepsilon_j^d \mid j \in [\![N_d]\!]\}$ for the current segment, and matches each true event to a default event window if their intersection-over-union (IoU) is at least 0.5.

At test time, we generate predictions across the default event windows and use a non-maximum suppression procedure to select between the candidate predictions. For a given class k, the procedure is as follows: First, the predictions are sorted according to probability of the event, which is above a threshold θ_k. Then, using the most probable prediction as an anchor, we sequentially evaluate the IoU between the anchor and the remaining candidate predictions, removing those with IoU larger than 0.5.

The output of the model is thus the set of predicted events $\boldsymbol{\varepsilon}^p = \{\mathbf{p}, \mathbf{y}\}$ containing the predicted class probabilities along with the corresponding onsets and durations

C. Signal Conditioning

We resampled all signals to a common sampling frequency of $f_s = 128\,\text{Hz}$ using a poly-phase filtering approach (Kaiser window, $\beta = 5.0$). Based on recommended filter specifications from the AASM, we designed Butterworth IIR filters for four sets of signals [1]. EEG and EOG channels were filtered using a 2nd order filter with a $0.3\,\text{Hz}$ to $35\,\text{Hz}$ passband, while chin and leg EMG channels were filtered using a 4th order high-pass filter with a $10\,\text{Hz}$ cut-off frequency. Nasal pressure channels was filtered using a 4th order high-pass filter with a $0.03\,\text{Hz}$ cut-off frequency, while thoracoabdominal channels were filtered using a 2nd order filter with a $0.1\,\text{Hz}$ to $15\,\text{Hz}$ passband. All filters were implemented using the zero-phase method.

Filtered signals were subsequently standardized by subtracting signal means and dividing by signal standard deviations for each PSG.

D. Target Encoding

For each data segment, target event classes $\boldsymbol{\pi} \in \mathbb{R}^{N_m\times K}$ were generated by one-hot encoding, and the target detection variable containing the onset and duration times $\mathbf{t} \in \mathbb{R}^{N_m\times 2}$ was encoded as

$$t_i = \left(\frac{\varrho_i^m - \varrho_j^d}{\delta_j^d}, \log \frac{\delta_i^m}{\delta_j^d} \right), \; i \in [\![N_m]\!], \; j \in [\![N_d]\!], \quad (1)$$

where ϱ_i^m is the center point of the true event matched to a default event window ϱ_j^d, and δ_i^m and δ_j^d are the corresponding durations of the true and default events.

E. Data Sampling

As the total number of default event windows N_d in a data segment most likely will be much higher than the number of event windows matched to a true event, i. e. $N_d \gg N_m$, we implemented a similar random data sampling strategy as in [28]. At training step t, a given PSGs record r has a certain number of associated number of Ar, LM, and SDB events ($n_{\text{Ar}}, n_{\text{LM}}, n_{\text{SDB}}$, respectively). We randomly sample a class k with equal probability $p_k = \frac{1}{K-1}$, while disregarding the negative class since this is most likely over-represented in the data segment in any case. Given the class k, we randomly sample an event ε_k from the PSG record r. We then randomly sample a $C \times T$ data segment with start time in the range $[\bar{\varepsilon}_k - T, \bar{\varepsilon}_k]$ where $\bar{\varepsilon}_k$ is the sample midpoint of the event ε_k. This ensures that a sampled data segment will contain at least $50\,\%$ of at least one event. We found that this approach to sampling data segments with a large ratio of negative to positive samples to be beneficial in all our experiments when monitoring the loss on the validation set.

F. Network Architecture

Similar to the architecture described in [30], we designed a splitstream network architecture, where each stream is responsible for the bulk feature extraction for a specific event class. For the given problem of detecting Ars, LMs, and SDBs, the network contains three streams: the Ar stream takes as input the EEGs, the EOGs, and the chin EMG signals for a total of $C_{\text{Ar}} = 5$ channels; the LM stream receives the $C_{\text{LM}} = 2$ leg EMG signals; and the SDB stream receives the nasal pressure and the thoracoabdominal signals for a total of $C_{\text{SDB}} = 3$ channels. An overview of the network architecture is shown graphically in Fig. 2.

1) Stream Specifics: Each stream is comprised of two components. First, a mixing module $\varphi_{\text{mix}} : \mathbb{R}^{C_*\times T} \to \mathbb{R}^{C_*\times T}$ computes a non-linear mixing of the C channels using a set of C single-strided 1-dimensional filters $\mathbf{w} \in \mathbb{R}^{C\times C}$ and rectified linear unit (ReLU) activation [45]. Second, the output activations from φ_{mix} are used as input to a feature extraction module $\varphi_{\text{feat}} : \mathbb{R}^{C_*\times T} \to \mathbb{R}^{f'\times T'}$, which transforms the input feature maps to a $f' \times T'$ feature space with a temporal dimension reduced by a factor of $\frac{T}{T'}$. The feature extraction module φ_{feat} is realized using $k_{\max}$ successive convolutions with an increasing number of filters $f' = f_0 2^{k-1}$, $k \in [\![k_{\max}]\!]$, where f_0 is a tunable base filter number. Each feature map is normalized using batch normalization [46] with subsequent ReLU activation.

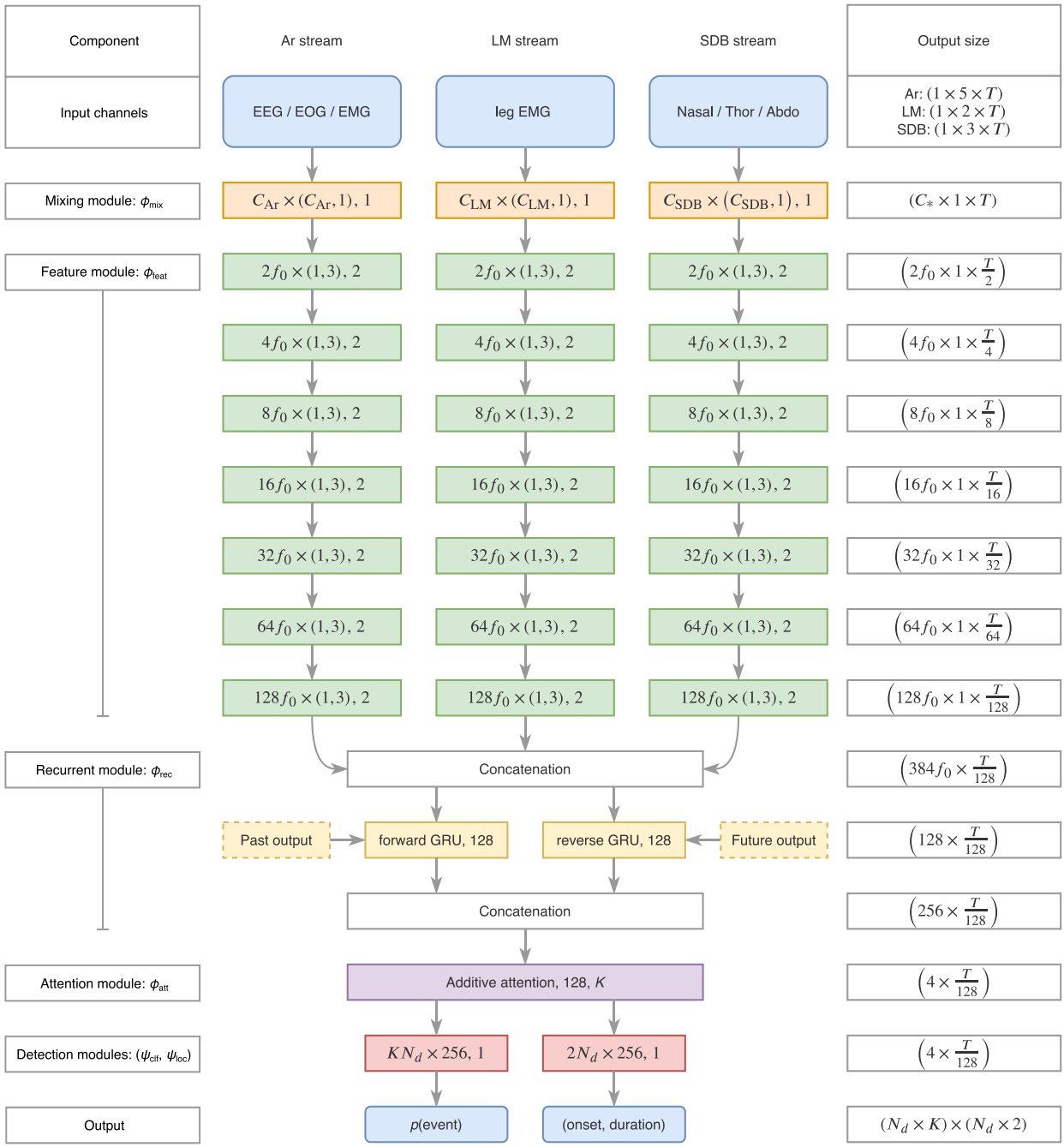

Fig. 2. MSED network architecture. Left column contains component names, while right column shows the output dimensions for each operation as (number of filters[x singleton] x time steps). Each stream in the middle (green) processes a separate set of input channels (blue, top), the results of which are concatenated before the bGRU (yellow). Outputs from the additive attention layer (purple) are convolved in the final classification and localization layers (red) to output the probabilities for each event class, and the predicted onset and duration of each event (blue, bottom). Convolution layers (orange, green, red) are detailed as [number of feature maps x kernel size, stride]. Recurrent layer (yellow) shows the direction and number of hidden units. Additive attention layer (purple) is described with the number of hidden and output units. GRU: gated recurrent unit; Ar: arousal; LM: leg movement; SDB: sleep disordered breathing.

2) Feature Fusion for Sequential Processing: The output vectors from each stream is concatenated into a combined feature vector $\mathbf{z} = (\mathbf{z}_{\text{ar}}, \mathbf{z}_{\text{lm}}, \mathbf{z}_{\text{sdb}}) \in \mathbb{R}^{3f' \times T'}$. We introduce sequential modeling of the feature vectors via a recurrent module $\varphi_{\text{rec}} : \mathbb{R}^{3f' \times T'} \to \mathbb{R}^{2n_h \times T'}$ realized with a bidirectional gated recurrent unit (bGRU) [47]. The output of the bGRU for timestep t is a vector $\mathbf{h}_t = (\mathbf{h}_t^{\text{f}}, \mathbf{h}_t^{\text{b}}) \in \mathbb{R}^{2n_h}$ containing the concatenated outputs from the forward (f) and backward (b) directions.

3) Additive Attention: We implemented a simple *additive attention* mechanism [48], which computes *context*-vectors $\mathbf{c} \in$

$\mathbb{R}^{2n_h \times K}$ for each event class as the weighted sum of the feature vector outputs $\mathbf{h} \in \mathbb{R}^{2n_h \times T'}$ from the bGRU.

Formally, attention is computed as

$$\mathbf{c} = \sum_{t}^{T'} \mathbf{h}_t \boldsymbol{\alpha}_t^{\intercal}, \tag{2}$$

where $\mathbf{h}_t$ is the feature vector for time step t, and $\boldsymbol{\alpha}_t \in \mathbb{R}^K$ is the attention weight computed as

$$\boldsymbol{\alpha}_t = \frac{\exp(\mathbf{W}_a \tanh(\mathbf{W}_u \mathbf{h}_t))}{\sum_{\tau}^{T'} \exp(\mathbf{W}_a \tanh(\mathbf{W}_u \mathbf{h}_\tau))}. \tag{3}$$

Here, $\mathbf{W}_u \in \mathbb{R}^{n_a \times 2n_h}$ and $\mathbf{W}_a \in \mathbb{R}^{K \times n_a}$ are learnable linear mappings of the feature vectors.

4) Detection: The final event classification and localization is handled by two modules, $\psi_{\text{clf}} : \mathbb{R}^{2n_h \times K} \to \mathbb{R}^{N_d \times K}$ and $\psi_{\text{loc}} : \mathbb{R}^{2n_h \times K} \to \mathbb{R}^{N_d \times 2}$, respectively. The classification module $\psi_{\text{clf}} : \mathbf{c} \mapsto \mathbf{p}$ outputs a tensor $\mathbf{p} \in [0,1]_+^{N_d \times K}$ containing predicted event class probabilities for each default event window. The localization module $\psi_{\text{loc}} : \mathbf{c} \mapsto \mathbf{y}$ outputs a tensor $\mathbf{y} \in \mathbb{R}^{N_d \times 2}$ containing encoded relative onsets and durations for a detected event for each default event window.

G. Training Objective

Similar to [37], we optimized the network parameters according to a three-component loss function consisting of: i) a localization loss ℓ_{loc}, ii) a positive classification loss ℓ_+, and iii) a negative classification loss ℓ_-, such that the total loss ℓ was defined by

$$\ell = \ell_{\text{loc}} + \ell_+ + \ell_-. \tag{4}$$

The localization loss ℓ_{loc} was calculated using a Huber function

$$\ell_{\text{loc}} = \frac{1}{N_+} \sum_{i \in \pi_+} f_H^{(i)}, \tag{5}$$

$$\mathbf{f}_H = \begin{cases} 0.5(\mathbf{y} - \mathbf{t})^2, & \text{if } |\mathbf{y} - \mathbf{t}| < 1, \\ |\mathbf{y} - \mathbf{t}| - 0.5, & \text{otherwise}, \end{cases} \tag{6}$$

where $i \in \pi_+$ yields indices of event windows with positive targets, i. e. event windows matched to an arousal, LM or SDB target, and N_+ is the number of positive targets in the given data segment.

The positive classification loss component ℓ_+ was calculated using a simple cross-entropy over the event windows matched to an arousal, LM, or SDB event:

$$\ell_+ = \frac{1}{N_+} \sum_{i \in \pi_+} \sum_{k \in [\![K]\!]} \pi_k^{(i)} \log p_k^{(i)}, \tag{7}$$

$$p_k^{(i)} = \frac{\exp s_k^{(i)}}{\sum_j \exp s_j^{(i)}}, \tag{8}$$

and $\pi_k^{(i)}$, $p_k^{(i)}$, and $s_k^{(i)}$ are the true class probability, predicted class probability, and logit score for the ith event window containing a positive sample.

Similar to [36], [49], the negative classification loss ℓ_- was calculated using a hard negative mining approach to balance the number of positive and negative samples in a data segment after matching default event windows to true events [50]. Specifically, this is accomplished by calculating the probability for the negative class (no event) for each unmatched default event window, and then calculating the cross entropy loss using the Z most probable samples. In our experiments, we set the ratio of positive to negative samples as 1:3, such that the calculation of ℓ involves $Z = 3$ times as many negative as positive samples.

We also explored a focal loss objective function for computing ℓ_+ and ℓ_- [51], however, we found that this approach severely deteriorated the ability of the network to accurately detect LM and SDB events compared to using worst negative mining.

TABLE II
MSED PARAMETER SETTINGS

Symbol	Description	Value
n_{train}	Number of train records	1653
n_{eval}	Number of eval records	200
n_{test}	Number of test records	1000
C	Number of input channels	10
δ	Duration of default event windows	3, 5, 10 s
N_d	Number of default event windows	208
f_s	Sampling rate	128 Hz
T	Segment length, samples	15360
T'	Reduced segment length, samples	120
$k_{\max}$	Number of layers in φ_{feat}	7
f_0	Base filter number	4
f'	Feature vector size	512
n_h	Hidden units in bGRU	128
n_u	Hidden units in φ_{att}	128
K	Number of output classes	4
(β_1, β_2)	Adam decay rates	$(0.9, 0.999)$
η	Initial learning rate	1×10^{-3}
	Learning rate decay factor	0.1
	Learning rate decay step	3 epochs
	Early stopping	10 epochs

H. Experimental Setups

In our experiments, we optimized the training objective using adaptive moment estimation (Adam) [52], according to the loss function described in Eq. (4). Exponential decay rates were fixed at $(\beta_1, \beta_2) = (0.9, 0.999)$, the learning rate at $\eta = 10^{-3}$, and $\epsilon = 10^{-8}$. The learning rate was decayed in a step-wise manner by multiplying η with a factor of 0.1 after 3 consecutive epochs with no improvement in loss value on the validation dataset. Similarly, we employed an early stopping scheme by monitoring the loss on the validation dataset and stopping the model training after 10 epochs of no improvement on $\mathcal{D}_{\text{EVAL}}$.

We tested four types of models in two categories: the first is a default split-stream model as shown in Fig. 2 with and without weight decay (splitstream, splitstream-wd). The second is a variation of the split-stream model, but where the ψ_{clf} and ψ_{loc} modules are realized using depth-wise convolutions, such that each attention group is used only for that type of event. The second category is also tested with and without weight decay (splitstream-dw, splitstream-dw-wd).

We benchmarked our proposed MSED model against DOSED by comparing overall performance on $\mathcal{D}_{\text{test}}$ after training on

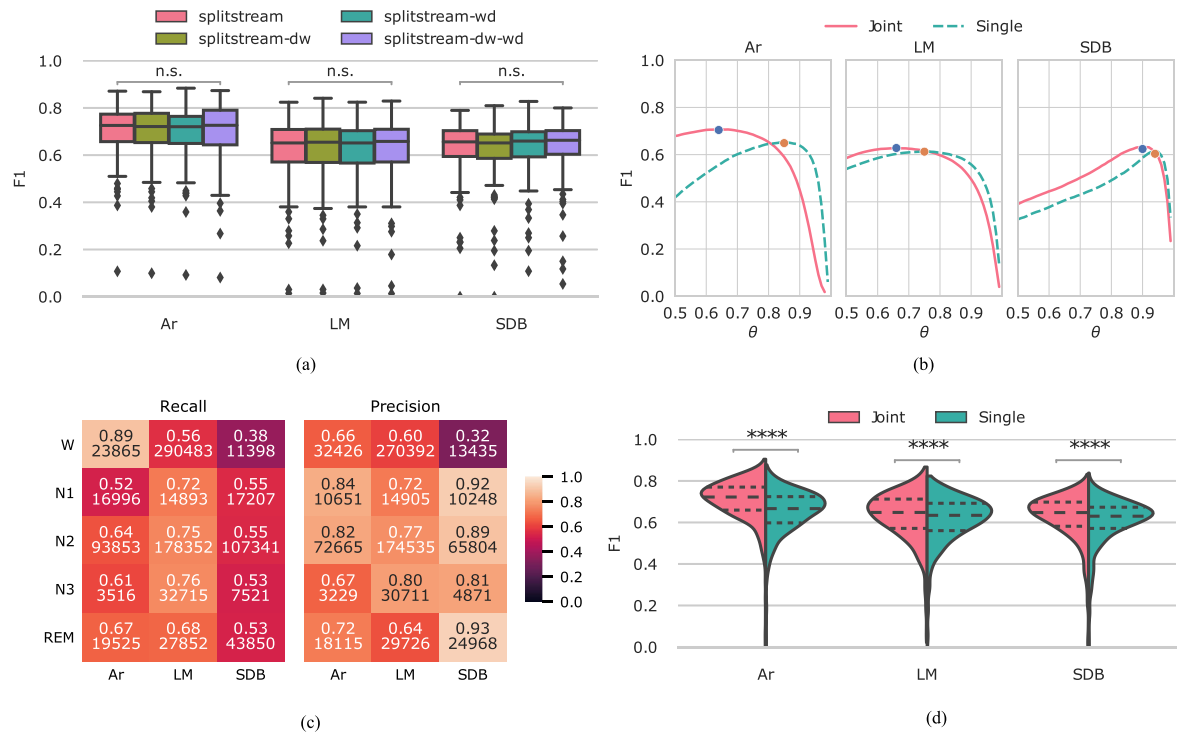

Fig. 3. Optimizing MSED architecture and model evaluation. **(a)** No significant differences in F1 were found between model architectures. **(b)** Optimizing F1 performance on $\mathcal{D}_{\text{eval}}$ as a function of detection threshold θ for joint (solid) and single-event (dashed) detection models. Blue and orange dots correspond to optimal θ/F1 pair. **(c)** Analysing event recall and precision categorized by sleep stage origin on $\mathcal{D}_{\text{test}}$ data. E.g. out of 19525 Ar events manually scored in REM, the optimal model correctly identifies 0.67%, while out of 10248 SDB events predicted in N1, 0.92% of these are also manually scored. **(d)** Evaluating optimized joint and single-event detection models on $\mathcal{D}_{\text{test}}$ given the respective detection threshold θ from **(b)**. For all three event types, the joint detection model outperforms the single-event models based on F1. Dashed lines in the violin plot interior show the associated 25th, 50th, and 75th percentiles. ****: $p < 0.001$. W: wakefulness; N1: non-REM stage 1; N2: non-REM stage 2; N3: non-REM stage 3; REM: rapid eye movement; Ar: arousal; LM: leg movement; SDB: sleep disordered breathing.

$n_{\text{train}} = 200$ subject PSGs. Each model was allowed 100 epochs of training, and the optimal models were selected based on lowest loss score on $\mathcal{D}_{\text{eval}}$ across epochs. F1, precision, and recall scores were obtained by evaluating optimized models on $\mathcal{D}_{\text{test}}$.

Various model parameters are shown in Table II.

I. Performance Evaluation

Performance was quantified using precision, recall, and F1 scores. Statistical significance in F1 score between groups was assessed with Kruskall-Wallis H-tests. The performance of joint vs. single-event detection models was tested with Wilcoxon signed rank tests for matched samples. The relationships between true and predicted arousal indices (ArIs), AHIs, and LMI were assessed using linear models and Pearson's r^2. Significance was set at $\alpha = 0.05$.

IV. Results and Discussion

A. Model Architecture Evaluation

We found no significant differences in F1 performance for either Ar (Kruskal-Wallis $H = 0.96$, $p = 0.81$), LM ($H = 0.23$, $p = 0.97$), or SDB detection ($H = 2.84$, $p = 0.42$), when evaluating the model architectures on $\mathcal{D}_{\text{eval}}$ (see Fig. 3(a)). Subsequent results are thus reported based on the default splitstream architecture.

B. Optimizing Threshold for Joint Vs. Single Event Detection

For each event type, we evaluated the F1 score as a function of classification threshold θ on $\mathcal{D}_{\text{eval}}$ for both the joint detection model as well as the single-event models. It can be observed in Fig. 3(b) that for all three events, the joint detection model achieves higher F1 score, although the increase is not as large for LM and SDB detection. This was also observed when evaluating the joint and single detection models with optimized thresholds on $\mathcal{D}_{\text{test}}$ for both Ar (Wilcoxon $W = 30440$, $p < 0.001$), LM ($W = 101103$, $p < 0.001$), and SDB detection ($W = 125461$, $p < 0.001$), see Fig. 3(d). Precision, recall and F1 scores for optimized models evaluated on are shown in Table IV. These findings provide evidence that the presence of certain event types can modulate the detection of other event types, and that this can be modeled using automatic methods. This is in line with what previous studies have found e. g. on event-by-event scoring agreement in arousals, which improved significantly from 58.7 % to 90.5 %, when including respiratory signals in the analysis [17].

TABLE III
TEMPORAL DIFFERENCE METRICS ACROSS EVENT TYPES AND PSGS FOR JOINT AND SINGLE PREDICTION MODELS

		Δduration	Δonset	Δoffset
Ar	Joint	4.81 ± 2.92 [1.07, 10.77]	-1.37 ± 1.39 [−3.99, 0.33]	3.44 ± 2.24 [0.56, 7.41]
	Single	6.42 ± 3.70 [1.37, 12.99]	-1.94 ± 1.93 [−5.33, 0.50]	4.49 ± 2.69 [0.85, 9.21]
LM	Joint	-0.46 ± 0.36 [−1.05, 0.07]	0.24 ± 0.22 [−0.04, 0.56]	-0.23 ± 0.20 [−0.55, 0.05]
	Single	-0.51 ± 0.39 [−1.16, 0.01]	0.26 ± 0.23 [−0.02, 0.61]	-0.25 ± 0.21 [−0.61, 0.06]
SDB	Joint	20.41 ± 13.36 [3.50, 43.77]	-9.39 ± 6.74 [−21.24, −0.97]	11.02 ± 7.67 [1.14, 23.70]
	Single	34.95 ± 16.63 [10.81, 63.18]	-16.00 ± 8.80 [−32.35, −3.65]	18.95 ± 9.51 [5.44, 35.33]

Positive values of onset/offset correspond to delayed predictions relative to the matched true event, while positive duration values correspond to an overestimation of event duration. Metrics are shown as mean ± standard deviation with 95% confidence interval in brackets across all events for all PSG recordings. Ar: arousal; LM: leg movement; SDB: sleep disordered breathing.

TABLE IV
PERFORMANCE SCORES FOR OPTIMIZED MODELS EVALUATED ON $\mathcal{D}_{\text{TEST}}$

		Joint	Single
Ar	F1	0.704 ± 0.106	0.649 ± 0.113
	Pr	0.759 ± 0.114	0.777 ± 0.107
	Re	0.672 ± 0.125	0.571 ± 0.127
LM	F1	0.628 ± 0.123	0.613 ± 0.116
	Pr	0.650 ± 0.169	0.661 ± 0.166
	Re	0.647 ± 0.120	0.607 ± 0.116
SDB	F1	0.624 ± 0.115	0.604 ± 0.118
	Pr	0.817 ± 0.142	0.835 ± 0.136
	Re	0.526 ± 0.146	0.492 ± 0.145
Overall	F1	0.652 ± 0.121	0.622 ± 0.117
	Pr	0.742 ± 0.159	0.758 ± 0.157
	Re	0.615 ± 0.146	0.556 ± 0.138

Metrics are shown aggregated across PSGs as mean ± standard deviation. Overall metrics are computed by averaging scores for each subject PSG. Ar: arousal; LM: leg movement; SDB: sleep disordered breathing; Pr: precision; Re: recall.

TABLE V
COMPARING MODEL PERFORMANCE AGAINST DOSED ON A REDUCED $\mathcal{D}_{\text{TRAIN}}$

		DOSED	MSED
Ar	F1	0.668 ± 0.115	0.677 ± 0.107
	Pr	0.711 ± 0.120	0.688 ± 0.118
	Re	0.651 ± 0.141	0.686 ± 0.131
LM	F1	0.619 ± 0.125	0.599 ± 0.127
	PR	0.634 ± 0.169	0.655 ± 0.174
	Re	0.644 ± 0.120	0.587 ± 0.133
SDB	F1	0.503 ± 0.123	0.626 ± 0.125
	Pr	0.706 ± 0.173	0.793 ± 0.146
	Re	0.412 ± 0.133	0.542 ± 0.157
Overall	F1	0.596 ± 0.140	0.634 ± 0.124
	Pr	0.684 ± 0.160	0.712 ± 0.159
	Re	0.568 ± 0.172	0.605 ± 0.153
No. parameters		385,489,502	9,435,606

Metrics are shown aggregated across PSGs as mean ± standard deviation. Overall metrics are computed by averaging scores for each subject PSG. Ar: arousal; LM: leg movement; SDB: sleep disordered breathing; Pr: precision; Re: recall.

C. Comparison With State-of-The-Art Multi-Event Detection

F1, recall, and precision scores for optimized DOSED and MSED models evaluated on $\mathcal{D}_{\text{test}}$ are shown in Table V. We observed an overall MSED F1 score of 0.634 ± 0.124 against an overall DOSED F1 score of 0.596 ± 0.140; and overall F1 scores for Ar, LM, and SDB of 0.677 ± 0.107, 0.599 ± 0.127, and 0.626 ± 0.125 for MSED, and 0.668 ± 0.115, 0.619 ± 0.125, and 0.503 ± 0.123 for DOSED. Factoring in the overall reduction in model size from 385,489,502 to 9,435,606 parameters, these results show the advantage of MSED compared to DOSED on the same dataset.

Comparing with existing single-event arousal detection models, MSED does not perform on the level of previous state-of-the-art proposed by Brink-Kjaer et al. [30]. Here, an EEG+EOG+EMG combination model for combined sleep-wake classification and arousal detection yielded an F1 score of 0.76, although this was reported on a much smaller dataset. Similarly, in the work by Carvelli et al. [31], a model combining two leg EMG channels achieved an impressive F1 score of 0.77, although this was also reported in a much smaller dataset. We did not perform in-depth ablations in this study, so it is possible that the MSED performance could be higher given sufficient fine-tuning. However, it is also worth noting that both of these models apply post-processing of the model outputs, most notably the removal of arousals and leg movements scored in wake, which is not performed in the current work, and fusion of events within certain manually-set thresholds. It is possible that this approach introduces a negative bias in our proposed model, since it is evident from Fig. 3(c) that the precision for all scored events is lower in W than in other sleep stages. In this work we wished the predictions to be orthogonal to manual sleep scoring, but future work should consider adding an automatic sleep scoring module to the model architecture.

While literature on sleep apnea detection is extensive, it is difficult to compare directly to the proposed approach, because the majority of studies either focus on obstructive apnea alone, do not report F1, precision, or recall scores, or only focus on the prediction of AHIs alone [53].

However, one recent study compared the event-by-event detection performance against a concensus score of five technicians. They reported an average human performance quantified by F1 of 0.55, and an F1 score from the automatic method of 0.57 [54]. Similarly, Nassi et al. [55] recently proposed their WaveNet model for precisely annotating SDB events in 1 s bins. Although their model also included post-processing of the bins, they obtained a mean F1 score across events of 0.406. We therefore see a marked improvement from the state of the art event detectors compared to MSED.

These results also indicate the massive research potential in terms of other ways to assess SDB; apart from AHI, which is

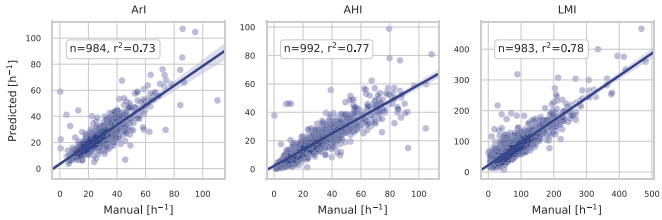

Fig. 4. Pearson correlation plots for each event type index between true and predicted values. The linear relationship is indicated with solid blue with 95% confidence intervals in light blue. ArI: arousal index; AHI: apnea-hypopnea index; LMI: leg movement index.

an average across the entire night, researchers and clinicians could potentially benefit from taking a more fine-grained approach. As illustrated by Chen et al. [56], patients with the exact same AHI can exhibit wildly different activity patterns (breathing disturbances), yet this is unaccounted for in state-of-the-art apnea detection models, as the majority of these are epoch-based [53]. Chen et al. propose "instantaneous AHI" as a solution to this problem, although their results were based on human annotation and not automatic detection as proposed here. The integration of these two methods would be an interesting approach for future research.

In addition, future work should explore novel methods for object detection in the computer vision literature. Most notably, the use of default event windows impose certain restrictions on the temporal scale of detected events, and this could be eliminated by using a Transformer-based detection model, such as the one found in the Detection Transformer [57]. Here, predictions are generated using a number of object queries, which are independent from the default event windows and thus not restricting the temporal scale of the events.

D. Correlation Between Experts and Model

For each event type, we computed the correlation coefficient between predicted and expert-annotated ArI, AHI, and LMI, which is shown in Fig. 4. We found a large positive correlation between true and predicted values for ArIs ($r^2 = 0.73$, $p < 0.001$), AHIs ($r^2 = 0.77, p < 0.001$), and LMI ($r^2 = 0.78$, $p < 0.001$).

A similar study by Biswal et al. [24] using an automatic method for automatic detection of SDB and LM events found similar or higher correlations between automatic and manual scoring ($r^2_{\text{SDB}} = 0.85$, and $r^2_{\text{LM}} = 0.79$, respectively), although their findings were based on almost five times as much data. Furthermore, obstructive, central, mixed and hypopneas with an associated 4% desaturation were lumped together into a single *apnea* class, which may have introduced unwanted bias towards obstructive apneas and hypopneas in their findings, since these are in general more prevalent than central and mixed apneas.

E. Temporal Difference Metrics

We compared the temporal precision between manual and MSED event scoring by looking at the errors in onset (Δonset), offsets (Δ offset), and durations (Δduration) calculated as

$$\Delta\,\text{onset} = \text{onset}_{\text{MSED}} - \text{onset}_{\text{manual}} \tag{9}$$

$$\Delta\,\text{offset} = \text{offset}_{\text{MSED}} - \text{offset}_{\text{manual}} \tag{10}$$

$$\Delta\,\text{duration} = \text{duration}_{\text{MSED}} - \text{duration}_{\text{manual}} \tag{11}$$

so that positive values of Δ onset, Δ offset corresponds to a positive shift to the right (delayed prediction), and positive values of Δ duration meaning an overestimation of the event duration compared to manual scoring.

Described in Table III, the model overestimates the duration of Ar events by a couple of seconds, which is caused by an earlier prediction of onset and delayed prediction of termination. For LM events, the model underestimates the duration by about half a second on average, which is due to earlier prediction of termination. For SDB events, the model overestimates the duration by about 25 seconds on average, which is caused by an earlier prediction of onset and delayed prediction of termination. These errors in predicted duration reflect the temporal characteristics of these events; LMs are shorter events (between 0.5 s to 10 s per definition), and it is thus unlikely to be overestimated by several seconds, while SDBs are longer events by one to two orders of magnitude, which also increases the size of the errors. Arousals are intermediate in length compared to LMs and SDBs, which is reflected in the error distributions.

V. Conclusion

We have presented a novel method for detecting short and long events present in polysomnogram recordings based on deep neural networks. Our method was able to distinguish between arousals, limb movements, and sleep-disordered breathing events with F1 scores of 0.70, 0.63, and 0.62, respectively, and we furthermore found that jointly optimizing a model for all three events performed better than the respective models optimized for each specific event type.

We benchmarked our algorithm against previous state-of-the-art and report an overall increase in F1 score from 0.60 to 0.63 despite a 97.5% reduction in memory footprint.

Furthermore, clinically relevant derivatives (ArIs, AHIs, LMI) showed a high positive correlation with manually computed values indicating a high degree of agreement between our model and experts.

Future work should incorporate ides from the object detection in computer vision literature and investigate more complex models with increased flexibility towards adding prediction capabilities for additional event types. Additionally, the low precision across all events observed during wakefulness could be remedied by incorporating an automatic sleep stage classification model which also merits further investigation.

Acknowledgment

The National Heart, Lung, and Blood Institute provided funding for the ancillary MrOS Sleep Study, "Outcomes of Sleep Disorders in Older Men," under the following grant numbers: R01 HL071194, R01 HL070848, R01 HL070847, R01 HL070842, R01 HL070841, R01 HL070837, R01 HL070838,

and R01 HL070839. The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

Some of the computing for this project was performed on the Sherlock cluster. We would like to thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to these research results.

The authors would like to thank Andreas Brink-Kjær, Rasmus Malik Thaarup Høegh, Anders Stevnhoved Olsen, Mads Olsen, and Laura Rose for their valuable comments and feedback in preparing this manuscript.

References

[1] R. B. Berry et al., *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Version 2.6*. Darien, IL, USA: American Academy of Sleep Medicine, 2020.

[2] P. Halász et al., "The nature of arousal in sleep," *J. Sleep Res.*, vol. 13, no. 1, pp. 1–23, doi: 10.1111/j.1365-2869.2004.00388.x.

[3] R. Ferri et al., "World association of sleep medicine (WASM) 2016 standards for recording and scoring leg movements in polysomnograms developed by a joint task force from the international and the European restless legs syndrome study groups(IRLSSG and EURLSSG)," *Sleep Med.*, vol. 26, pp. 86–95, 2016, doi: 10.1016/j.sleep.2016.10.010.

[4] R. Ferri et al., "Periodic leg movements during sleep: Phenotype, neurophysiology, and clinical significance," *Sleep Med.*, vol. 31, pp. 29–38, 2017, doi: 10.1016/j.sleep.2016.05.014.

[5] R. S. Rosenberg and S. Van Hout, "The american academy of sleep medicine inter-scorer reliability program: Respiratory events," *J. Clin. Sleep Med.*, vol. 10, no. 4, pp. 447–454, 2014, doi: 10.5664/jcsm.3630.

[6] R. G. Norman et al., "Interobserver agreement among sleep scorers from different centers in a large dataset," *Sleep*, vol. 23, no. 7, pp. 1–8, 2000, doi: 10.1093/sleep/23.7.1e.

[7] H. Danker-Hopfe et al., "Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders," *J. Sleep Res.*, vol. 13, pp. 63–69, 2004, doi: 10.1046/j.1365-2869.2003.00375.x.

[8] H. Danker-Hopfe et al., "Interrater reliability for sleep scoring according to the rechtschaffen & kales and the new AASM standard," *J. Sleep Res.*, vol. 18, no. 1, pp. 74–84, 2009, doi: 10.1111/j.1365-2869.2008.00700.x.

[9] R. S. Rosenberg and S. Van Hout, "The american academy of sleep medicine inter-scorer reliability program: Sleep stage scoring," *J. Clin. Sleep Med.*, vol. 9, pp. 81–87, 2013, doi: 10.5664/jcsm.2350.

[10] X. Zhang et al., "Process and outcome for international reliability in sleep scoring," *Sleep Breathing*, vol. 19, no. 1, pp. 191–195, 2015, doi: 10.1007/s11325-014-0990-0.

[11] M. Younes, J. Raneri, and P. Hanly, "Staging sleep in polysomnograms: Analysis of inter-scorer variability," *J. Clin. Sleep Med.*, vol. 12, no. 6, pp. 885–894, 2016, doi: 10.5664/jcsm.5894.

[12] M. Younes et al., "Reliability of the American academy of sleep medicine rules for assessing sleep depth in clinical practice," *J. Clin. Sleep Med.*, vol. 14, no. 2, pp. 205–213, 2018, doi: 10.5664/jcsm.6934.

[13] M. J. Drinnan et al., "Interobserver variability in recognizing arousal in respiratory sleep disorders," *Amer. J. Respir. Crit. Care Med.*, vol. 158, pp. 358–362, 1998, doi: 10.1164/ajrccm.158.2.9705035.

[14] C. W. Whitney et al., "Reliability of scoring respiratory disturbance indices and sleep staging," *Sleep*, vol. 21, no. 7, pp. 749–757, 1998, doi: 10.1093/sleep/21.7.749.

[15] J. S. Loredo et al., "Night-to-night arousal variability and interscorer reliability of arousal measurements," *Sleep*, vol. 22, no. 7, pp. 916–920, 1999, doi: 10.1093/sleep/22.7.916.

[16] M. Smurra et al., "Sleep fragmentation: Comparison of two definitions of short arousals during sleep in OSAS patients," *Eur. Respir. J.*, vol. 17, pp. 723–727, 2001, doi: 10.1183/09031936.01.17407230.

[17] R. J. Thomas, "Arousals in sleep-disordered breathing: Patterns and implications," *Sleep*, vol. 26, no. 8, pp. 1042–1047, 2003.

[18] M. H. Bonnet et al., "The scoring of arousal in sleep: Reliability, validity, and alternatives," *J. Clin. Sleep Med.*, vol. 3, no. 2, pp. 133–145, 2007, doi: 10.5664/jcsm.26815.

[19] U. J. Magalang et al., "Agreement in the scoring of respiratory events and sleep among international sleep centers," *Sleep*, vol. 36, no. 4, pp. 591–596, 2013, doi: 10.5665/sleep.2552.

[20] H. Koch et al., "Automatic sleep classification using a data-driven topic model reveals latent sleep states," *J. Neurosci. Methods*, vol. 235, pp. 130–137, 2014, doi: 10.1016/j.jneumeth.2014.07.002.

[21] A. Supratak et al., "DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 11, pp. 1998–2008, Nov. 2017, doi: 10.1109/TNSRE.2017.2721116..

[22] S. Chambon et al., "A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 4, pp. 758–769, Apr. 2018, doi: 10.1109/TNSRE.2018.2813138.

[23] A. N. Olesen et al., "Deep residual networks for automatic sleep stage classification of raw polysomnographic waveforms," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Soc., Honolulu, HI, USA: IEEE, 2018, pp. 1–4, doi: 10.1109/EMBC.2018.8513080.

[24] S. Biswal et al., "Expert-level sleep scoring with deep neural networks," *J. Amer. Med. Informat. Assoc.*, vol. 25, no. 12, pp. 1643–1650, 2018, doi: 10.1093/jamia/ocy131.

[25] J. B. Stephansen et al., "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy," *Nat. Commun.*, vol. 9, no. 1, 2018, Art. no. 5229, doi: 10.1038/s41467-018-07229-3.

[26] H. Phan et al., "Joint classification and prediction CNN framework for automatic sleep stage classification," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1285–1296, May 2019, doi: 10.1109/TBME.2018.2872652.

[27] H. Phan et al., "SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 3, pp. 400–410, Mar. 2019, doi: 10.1109/TNSRE.2019.2896659.

[28] A. N. Olesen et al., "Towards a flexible deep learning method for automatic detection of clinically relevant multi-modal events in the polysomnogram," in *Proc. IEEE 41st Annu. Int. Conf. Eng. Med. Biol. Soc.*, Berlin, Germany: IEEE, 2019, pp. 556–561, doi: 10.1109/EMBC.2019.8856570.

[29] D. Alvarez-Estevez and I. Fernández-Varela, "Large-scale validation of an automatic EEG arousal detection algorithm using different heterogeneous databases," *Sleep Med.*, vol. 57, pp. 6–14, 2019, doi: 10.1016/j.sleep.2019.01.025.

[30] A. Brink-Kjaer et al., "Automatic detection of cortical arousals in sleep and their contribution to daytime sleepiness," *Clin. Neurophysiol.* vol. 131, no. 6, pp. 1187–1203, 2020, doi: 10.1016/j.clinph.2020.02.027.

[31] L. Carvelli et al., "Design of a deep learning model for automatic scoring of periodic and non-periodic leg movements during sleep validated against multiple human experts," *Sleep Med.*, vol. 69, pp. 109–119, 2020, doi: 10.1016/j.sleep.2019.12.032.

[32] M. Ghassemi et al., "You snooze, you win: The PhysioNet/computing in cardiology challenge 2018," in *Proc. Comput. Cardiol. Conf.*, vol. 45, 2018, pp. 1–4, doi: 10.22489/CinC.2018.049.

[33] A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000, doi: 10.1161/01.CIR.101.23.e215.

[34] A. Brink-Kjaer et al., "Cortical arousal frequency is increased in narcolepsy type 1," *Sleep*, vol. 44, no. 5, 2021, Art. no. zsaa255, doi: 10.1093/sleep/zsaa255.

[35] A. Brink-Kjær et al., "Arousal characteristics in patients with Parkinson's disease and isolated rapid eye movement sleep behavior disorder," *Sleep*, vol. 44, no. 12, 2021, Art. no. zsab167, doi: 10.1093/sleep/zsab167.

[36] S. Chambon et al., "DOSED: A deep learning approach to detect multiple sleep micro-events in EEG signal," *J. Neurosci. Methods*, vol. 321, pp. 64–78, 2019, doi: 10.1016/j.jneumeth.2019.03.017.

[37] A. N. Olesen, "Deep transfer learning for improving single-EEG arousal detection," in *Proc. IEEE 42nd Annu. Int. Conf. Eng. Med. Biol. Soc.*, Montreal, QC, Canada: IEEE, 2020, pp. 99–103, doi: 10.1109/EMBC44109.2020.9176723.

[38] J. B. Blank et al., "Overview of recruitment for the osteoporotic fractures in men study (MrOS)," *Contemporary Clin. Trials*, vol. 26, no. 5, pp. 557–568, 2005, doi: 10.1016/j.cct.2005.05.005.

[39] E. Orwoll et al., "Design and baseline characteristics of the osteoporotic fractures in men (MrOS) study – A large observational study of the determinants of fracture in older men," *Contemporary Clin. Trials*, vol. 26, no. 5, pp. 569–585, 2005, doi: 10.1016/j.cct.2005.05.006.

[40] T. Blackwell et al., "Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: The osteoporotic fractures in men sleep study," *J. Amer. Geriatr. Soc.*, vol. 59, no. 12, pp. 2217–2225, 2011, doi: 10.1111/j.1532-5415.2011.03731.x.

[41] American Sleep Disorders Association, "EEG arousals: Scoring rules and examples: A preliminary report from the sleep disorders atlas task force of the American sleep disorders association," *Sleep*, vol. 15, no. 2, pp. 173–184, 1992, PMID: 11032543.

[42] M. Zucconi et al., "The official world association of sleep medicine (WASM) standards for recording and scoring periodic leg movements in sleep (PLMS) and wakefulness (PLMW) developed in collaboration with a task force from the international restless legs syndrome study group," *Sleep Med.*, vol. 7, no. 2, pp. 175–183, 2006, doi: 10.1016/j.sleep.2006.01.001.

[43] D. A. Dean et al., "Scaling up scientific discovery in sleep medicine: The national sleep research resource," *Sleep*, vol. 39, no. 5, pp. 1151–1164, 2016, doi: 10.5665/sleep.5774.

[44] G.-Q. Zhang et al., "The national sleep research resource: Towards a sleep data commons," *J. Amer. Med. Informat. Assoc.*, vol. 25, no. 10, pp. 1351–1358, 2018, doi: 10.1093/jamia/ocy064.

[45] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, Haifa, Israel, 2010. [Online]. Available: https://icml.cc/Conferences/2010/papers/432.pdf

[46] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, Lille, France: JMLR, 2015, *arXiv: 1502.03167 [cs.LG]*.

[47] K. Cho et al., "On the properties of neural machine translation: Encoder–decoder approaches," in *Proc. 8th Work Syntax Semantics Struct. Stat. Transl.*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2014, pp. 103–111, doi: 10.3115/v1/W14-4012.

[48] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent.*, San Diego, CA, USA, 2015, *arXiv: 1409.0473 [cs.CL]*.

[49] S. Chambon et al., "A deep learning architecture to detect events in EEG signals during sleep," in *Proc. IEEE 28th Int. Work. Mach. Learn. Signal Process.*, Aalborg, Denmark: IEEE, 2018, pp. 1–6, doi: 10.1109/MLSP.2018.8517067.

[50] W. Liu et al., "SSD: Single shot MultiBox detector," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[51] T.-Y. Lin et al., "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020, doi: 10.1109/TPAMI.2018.2858826.

[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv: 1412.6980 [cs.LG]*.

[53] F. Mendonça et al., "A review of obstructive sleep apnea detection approaches," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 825–837, Mar. 2019, doi: 10.1109/JBHI.2018.2823265.

[54] V. Thorey et al., "AI vs humans for the diagnosis of sleep apnea," in *Proc. IEEE 41st Annu. Int. Conf. Eng. Med. Biol. Soc.*, EMBC, Berlin, Germany: IEEE, 2019, pp. 1596–1600, doi: 10.1109/EMBC.2019.8856877.

[55] T. E. Nassi et al., "Automated scoring of respiratory events in sleep with a single effort belt and deep neural networks," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 6, pp. 2094–2104, Jun. 2022, doi: 10.1109/TBME.2021.3136753.

[56] S. Chen, "Dynamic models of obstructive sleep apnea provide robust prediction of respiratory event timing and a statistical framework for phenotype exploration," *Sleep*, vol. 45, no. 12, Art. no. zsac189, doi: 10.1093/sleep/zsac189.

[57] N. Carion et al., "End-to-end object detection with transformers," in *Proc. 16th Eur. Conf. Comput. Vis.*, A. Vedaldi et al., Eds., Lecture Notes in Computer Science, Cham: Springer, vol. 12346, 2020, pp. 213–229, doi: 10.1007/978-3-030-58452-8_13.