DistillSleep: Real-Time, On-Device, Interpretable Sleep Staging from Single-Channel EEG

Keondo Park[1], Joopyo Hong[1], Wooseok Lee[1], Hyun-Woo Shin[234], and Hyung-Sin Kim[1]

Seoul National University Graduate School, Seoul, Republic of Korea, Graduate School of Data Science [1]

Obstructive Upper Airway Research (OUaR) Laboratory, Department of Pharmacology, Seoul National University College of Medicine, Seoul, Republic of Korea [2]

OUaR LaB, Inc, Seoul, Republic of Korea[3]

Department of Otorhinolaryngology-Head and Neck Surgery, Seoul National University Hospital, Seoul, Republic of Korea[4]

Corresponding authors:

Hyung-Sin Kim, Graduate School of Data Science, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul (08826), Republic of Korea, Email: hyungkim@snu.ac.kr;

Hyun-Woo Shin, Department of Pharmacology, Seoul National University College of Medicine and Department of OtorhinolaryngologyHead and Neck Surgery, Seoul National University Hospital, 103 Daehak-ro, Jongno-gu, Seoul (03080), Republic of Korea. Email: charlie@snu.ac.kr

Abstract

Study objectives: Polysomnography (PSG) is the current gold standard for sleep staging but requires laboratory equipment, multiple sensors, and labor-intensive manual scoring. We developed *DistillSleep*, a single-channel EEG framework that delivers accurate, real-time, and interpretable sleep staging on resource-constrained devices.

Methods: *DistillSleep* consists of (1) a high-capacity teacher model and (2) a 109 k-parameter student model designed for edge deployment. Both incorporate a Multi-Wavelength Pyramid module and Transformer-based architecture to capture intra- and inter-epoch features. Feature- and prediction-level knowledge distillation transfers the teacher's expertise to the student. Training and evaluation used >10,000 overnight recordings from six cohorts (SHHS1, PhysioNet 2018, DCSM, KISS, Sleep-EDF, ISRUC), following AASM guidelines. Performance was assessed with Macro-F1.

Results: The teacher achieved state-of-the-art Macro-F1 scores (SHHS1 81.1%, PhysioNet 78.9%, DCSM 81.2%, KISS 80.0%) and provided frequency-resolved saliency maps, inter-epoch context and well-calibrated confidence (ECE 0.07). The student maintained competitive accuracy (up to 79.7% Macro-F1) while executing <10 ms per 30-second epoch on three embedded platforms (Raspberry Pi 4, Jetson orin nano, Coral dev board), reducing computational load 115-fold versus the best prior method (SleePyCo).
Interpretability was transferred intact to the student, offering clinicians frequency-band importance and inter-epoch context visualizations, and calibration was further improved by 2.7 X.

Conclusions: *DistillSleep* combines expert-level accuracy, millisecond-scale latency, and transparent decision logic in a single-channel EEG form factor. These capabilities pave the way for point-of-care diagnostics, same-night therapy titration, and large-scale home monitoring, expanding the reach of sleep medicine while retaining clinical trust.

Keywords
Automatic sleep staging, Knowledge distillation, Interpretable deep learning, On-device AI, Machine Learning, EEG analysis, Big data

Statement of Significance

Polysomnography is the gold standard for sleep staging, but its high cost, laboratory equipment, and lengthy manual scoring limit patient access. *DistillSleep* replaces the typical 12-20-sensor setup with a single-channel EEG and performs inference in <10 ms per epoch on a Raspberry Pi, Jetson orin nano, or Coral dev board. Trained and tested on >10,000 overnight recordings from six independent cohorts, it matches expert accuracy (Macro-F1 up to 80%) and supplies clinicians with frequency-band saliency, inter-epoch context, and well-calibrated confidence scores. By combining interpretability, millisecond-level latency, and an open-source code release, *DistillSleep* supports point-of-care diagnostics, same-night CPAP titration, and large-scale home monitoring, substantially broadening the reach of sleep medicine.

1. Introduction

Sleep disorders are common, costly, and dangerous. An estimated 936 million adults worldwide suffer from clinically significant sleep disorders [1], most of which remain undiagnosed or undertreated [2], costing five major OECD economies (US, UK, JPN, GER, CAN) approximately $680 billion annually [3].
Accurate sleep staging - classifying sleep into Wake, N1, N2, N3, and REM in 30-second epochs - is therefore critical for diagnosis and therapy titration [4,5].
However, the current gold standard, polysomnography (PSG), is expensive, intrusive, and capacity-limited. Patients must endure attachment to 12--20 sensors, while clinicians require 90--120 minutes per patient for manual scoring, achieving only approximately 82% inter-scorer agreement [6-8]. This explains why most of sleep disorder cases do not receive definitive diagnosis and timely treatment [peppard2013increased], exacerbating risks such as cardiovascular disease, daytime impairment, and increased accidents [9,10].
Real-time, portable, and interpretable sleep-staging systems can close this gap by enabling point-of-care decisions and reducing clinical workload. Continuous staging supports same-night continuous positive airway pressure (CPAP) titration and closed-loop ventilator control [11,12]. Automated home reports streamline patient follow-up, reducing manual scoring burden [13].
Our focus is to realize these benefits with a compact, interpretable single-channel electroencephalogram (EEG) model that performs inference in <10 ms and runs on bedside monitors or wearable headbands.
Longer-term opportunities, such as adaptive bedroom thermoregulation [14,15], adjustable beds [16], and memory-consolidation stimulation [17], become feasible once trustworthy real-time staging is widely available.
Rapid advances in deep learning have produced sleep staging models that approach expert accuracy on single-channel EEG [18-24].
However, most published models remain computationally heavy black boxes [18-22,24], requiring server-class GPUs, long inference times, and unclear decision logic, which are incompatible with bedside monitors, wearables, and clinicians' need for transparent evidence.
Balancing strong performance and interpretability within streamlined, on-device models thus remains under-explored [23]
In this study, we propose *DistillSleep*, a deep learning framework for on-device, real-time, and interpretable automatic sleep staging based on single-channel EEG. *DistillSleep* consists of a large, high-performing teacher model and an *exceptionally lightweight* student model.
The teacher model (*DistillSleep*-T) achieves state-of-the-art performance robustly in four large datasets, with Macro-F1 scores of 81.1% on SHHS1 [25,26], 78.9% on Physionet 2018 [27,28], 81.2% on DCSM, and 80.0% on KISS [29]. To address “black-box” skepticism, it provides comprehensive interpretability, including frequency-level importance, inter-epoch relationships, and well-calibrated confidence scores, ensuring both robust performance and clinical trustworthiness.
More importantly, the student model (*DistillSleep*-S) is designed for on-device real-time applications, featuring only 109k parameters (59.4 times smaller than the teacher model) and completing single-epoch inference within 10 ms on three resource-constrained platforms. Compared to SleePyCo [24], the best-performing conventional method, *DistillSleep*-S reduces model size by 20.8 times and computation overhead by 114.9 times. To achieve this efficiency without sacrificing performance, we design a knowledge distillation process [30-32] that enables *DistillSleep*-S to learn not only from the training data labels but also from the high-performing teacher model.
DistillSleep-S is validated on the four large public cohorts, and two additional independent datasets (SleepEDF-78 dataset (78 subjects) [27,33] and ISRUC dataset (118 subjects) [34]), >10,000 overnight studies. Despite its compact size, *DistillSleep*-S achieves competitive Macro-F1 scores up to 79.7% while maintaining comprehensive interpretability and improving confidence calibration.
Its real-time capability, reliability, and transparency have the potential to broaden access to sleep diagnostics and underpin emerging applications that depend on immediate sleep-stage feedback.
Our code and trained weights are open-sourced at: https://github.com/KeondoPark/sleep.git

2. Methods

DistillSleep comprises a high-performing, large-scale teacher model and an exceptionally compact student model.
Both models adopt a *hybrid two-stage architecture* that combines convolutional neural network (CNN) and Transformer. The CNN-based stage 1 model effectively captures *intra-epoch* features and predicts the sleep stage for each epoch (one-to-one). Subsequently, the Transformer-based stage 2 model incorporates *inter-epoch* features by taking a sequence of feature vectors from stage 1 model as input, to produce the refined sequence of sleep staging results (many-to-many). Moreover, our tailored architecture in both stages offers comprehensive interpretability.
While the teacher and student models share architectural similarities, they are designed for different purposes. The teacher model (*DistillSleep*-T) is designed to deliver precise and robust sleep stage classification results. In contrast, the student model (*DistillSleep*-S) focuses on reducing model size and computational requirements while maintaining effective performance. To achieve this significant size reduction without compromising performance and comprehensive interpretability, we employed knowledge distillation, a technique that transfers knowledge from the high-performing teacher model to the lightweight student model.

2.1 Model Architecture

2.1.1 Teacher Model

The overall architecture of teacher model is presented in Figure 1a. In the first stage, the model learns *intra-epoch*

patterns using a combination of convolutional blocks, a newly proposed Multi-Wavelength Pyramid (MWP) module, and inverted residual blocks [35]. The MWP module is particularly crucial in this stage, allowing the model to capture multi-scale patterns that correspond to the diverse frequency characteristics of each sleep stage. Specifically, each sleep stage is characterized by a predominant EEG frequency or wave types (Wake: above 7Hz (Alpha-Gamma), N1: 4-7Hz (Theta), N2: 4-7Hz (Theta), N3: 0.5-4Hz (Delta), REM: 2-6Hz (Theta-Delta)) [36-38]. To achieve accurate sleep staging, it is essential to extract features from these distinct EEG wavelength ranges.

To this end, as depicted in Figure 1c, the MWP module defines four fields of views (FoVs): (1) Supra-Alpha (Alpha, Beta, Gamma and shorter wavelengths), (2) Theta, (3) Delta, and (4) Infra-Delta (longer wavelengths beyond Delta). This is achieved through parallel max-pooling operations with varying pool sizes of 15, 25, 200, and 400, corresponding to different FoVs.

Concretely, the MWP module applies the four separate max-pooling operations to features extracted by the third convolutional block, then concatenates these pooled features to create a multi-scale representation. This representation is processed through a series of inverted residual blocks [35], which not only preserves important information but also enhances computational efficiency. In addition, the MWP module offers *intra-epoch* interpretability by analyzing the contribution of each max-pooled feature to the model's predictions.

The stage 2 model focuses on learning temporal dependencies across consecutive sleep epochs. To achieve this, this model takes the sequence of feature vectors from the stage 1 model as input, incorporating position encoding [39] to provide the sequence order. The hierarchical transformer blocks, inspired by Swin-transformer [40,41], constitute the core of the stage 2 model. In this hierarchical setup, lower transformer blocks employ attention locally, while higher transformer blocks use attention on broader temporal spans, allowing the model to progressively capture higher-level dependencies across the sequence. As presented in Figure 1d, each hierarchical level consists of two transformer blocks with identical local window sizes. In the first transformer block, the sequence is divided by local windows and attention is applied within each local window. However, this block alone cannot capture relationships between patches spanning neighboring windows. To overcome this limitation, the second block uses shifted windowing approach: the sequence is shifted by half the window size to create a new set of local windows, thereby enhancing the model's ability to capture more diverse local dependencies. After each pair of transformer blocks, two adjacent patches are merged into a single patch, making each patch have a higher-level feature representation. The model comprises four hierarchical levels, with the highest level ultimately applying attention across the entire sequence. Notably, the highest level omits the shifted transformer block and patch merging, because it has access to the complete sequence information. Consequently, the model includes seven transformer blocks in total. This hierarchical architecture facilitates the efficient integration of dependencies across sequential epochs, which is critical for accurate sleep staging. The output from the transformer blocks is then passed to a classifier, which generates the final probability distribution for each class. Additionally, we implement an ensemble technique, following previous works [22,23,42,43], to obtain the final prediction results based on multiple sequence predictions. Since the stage 2 model uses a sequence of epochs as input, each epoch appears at different positions within the sequence. For instance, if the sequence shifts by one epoch, the last epoch in the current sequence moves to the second-to-last position in the next sequence, and so forth. Consequently, each epoch is predicted multiple times, appearing in progressively earlier positions across different sequences. These overlapping predictions, which are probability vectors for each class, are then averaged to produce a final, more robust decision for each epoch, ultimately enhancing the model's accuracy and stability.

2.1.2 Student Model

The overall architecture of student model is presented in Figure 1b. The student model is a lightweight version of the teacher model, maintaining the same two-stage architecture while being optimized for enhanced efficiency. Similar to the teacher model, the student model incorporates convolutional blocks, MWP layers, and inverted residual blocks in stage 1, and followed by a transformer-based structure in stage 2. In stage 1, the student model utilizes fewer inverted residual blocks with smaller number of filters compared to the teacher model, allowing for a more compact design. In stage 2, whereas the teacher model employs hierarchically stacked seven transformer blocks, the student model simplifies this by using only a single transformer block with reduced dimension. The single transformer block relies solely on global attention, allowing it to capture dependencies across the entire sequence efficiently. To further reduce the computational burden, layer normalization and GELU activation in the teacher model are replaced by batch normalization and ReLU6 activation. This reduction in complexity makes the student model more suitable for resource-constrained environments while retaining the essential structure of the teacher model.

2.2 Loss

2.2.1 Weighted Cross Entropy Loss

In sleep datasets the class distribution is commonly imbalanced. In such cases, using the plain cross entropy loss for training commonly leads to the biased model towards majority classes. To tackle this, we used weighted cross entropy loss, similar to the ones used in previous studies [44,45]. More specifically, the weighted cross entropy (WCE) loss is defined as follows:

$$Loss_{WCE} = -\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N_k} w^k y_{(i,k)} \log \hat{y}_{(i,k)}$$

where k represents each sleep stage, N_k is the total

number of epochs labeled as the stage k from the training split, $N(=\sum_{k=1}^{K} N_k)$ is the total number of epochs from the training split, $y_{(i,k)}$ and $\hat{y}_{(i,k)}$ are the ground truth label and the predicted sleep stage by the model for i-th epoch in stage k, respectively, and $w^k = \max\left(\log_K \frac{N}{N_k}, 1\right)$ is the weight assigned to the loss corresponding to each sleep stage k. This loss was used solely to train *DistillSleep*-T and used in conjunction with knowledge distillation loss to train *DistillSleep*-S.

2.2.2 Knowledge Distillation Loss

The knowledge is transferred to *DistillSleep*-S by using knowledge distillation loss term along with weighted cross entropy loss. When the student model is trained to mimic the teacher model's intermediate representations, it is referred to as feature-based knowledge distillation. Conversely, when the student model is trained to replicate the teacher model's predictions, this process is referred to as logit-based knowledge distillation. For feature distillation in stage 1, we used OFD (A comprehensive Overhaul of Feature Distillation [31]). The features before MWP and the output features of penultimate layer were used for feature distillation in stage 1 model. The feature distillation (FD) loss is calculated as following:

$$Loss_{FD,s1} = \sum_{l \in L} w_l d_p \left(MReLU(F_l^T), r(F_l^S) \right)$$

where L represents the set of layers to be distilled, l is each layer in L and w_l is the weight given to each layer's distillation loss. MReLU is the marginal ReLU which is the modified ReLU having lower bound less than zero, $\boldsymbol{F}_l^{T,S}$ are the features of the teacher network and the student network, and r is the regressor which projects the student's features to teacher's features for dimension matching. The partial distance function d_p is defined as following:

$$d_p(\boldsymbol{T}, \boldsymbol{S}) = \Sigma_{i \in dim(\boldsymbol{T})} \begin{cases} 0 & if\ S_i \le T_i \le 0, \\ S_C(T_i, S_i) & otherwise \end{cases}$$

where S_C is cosine similarity.

For feature distillation in stage 2, we used CCKD (Correlation congruence for knowledge distillation [46]) and attention similarity loss [47]. Specifically, the FD loss is calculated as following:

$$Loss_{FD,s2} = \Sigma_{l \in L}[w_l MSE(\boldsymbol{F}_l^T(\boldsymbol{F}_l^T)^\top, r(\boldsymbol{F}_l^S) r(\boldsymbol{F}_l^S)^\top)] + w_A S_C(\boldsymbol{A}^T, \boldsymbol{A}^S)$$

where L represents the set of layers to be distilled, l is each layer in L and w_l is the weight given to each layer's distillation loss. $\boldsymbol{F}_l^{T,S}$ are the output features of the teacher and the student network from distilling layer, r is the regressor for dimension matching and MSE is mean squared error. $\boldsymbol{A}^{T,S}$ are the vectorized global attention, S_C is cosine similarity and w_A is the weight given to attention similarity loss.

We employed improved variant of DIST (Knowledge distillation from a stronger teacher [32]) for logit-based knowledge distillation. Our logit-distillation (LD) loss is decomposed into two terms: intra-class loss and inter-class loss. Intra-class loss aligns class-wise probabilities of the student across all data points in the batch with those of the teacher, ensuring consistency within each class. In contrast, inter-class loss encourages the student's output probability vector for each data point to closely match that of the teacher. Unlike the original DIST, we integrate the class-wise weights in both intra- and inter-class loss to account for class imbalance. Formally, the LD loss is defined as following:

$$Loss_{LD} = Loss_{Inter} + Loss_{Intra}$$

$$Loss_{Intra} = \frac{1}{K} \sum_{k=1}^{K} w^k d\left(\hat{y}_{(:,k)}^T, \hat{y}_{(:,k)}^S\right)$$

$$Loss_{Inter} = \frac{1}{N} \sum_{k=1}^{K} w_i d\left(\hat{y}_{(i,:)}^T, \hat{y}_{(i,:)}^S\right)$$

where $\hat{y}_{(i,k)}^{T,S}$ are the predicted probability for i-th data and k-th class, of the teacher/student models. The same class-wise weight in weighted cross entropy loss is applied as the weight for the intra-class loss, w^k. The weight for the inter-class loss, w_i follows w^k, based on each data point i's ground truth sleep stage k.

d is the distance function between two probability vectors from teacher and student, which is defined as following:

$$d(\boldsymbol{u}, \boldsymbol{v}) = 1 - \rho(\boldsymbol{u}, \boldsymbol{v})$$

where $\rho(\boldsymbol{u}, \boldsymbol{v})$ is Pearson correlation coefficient between two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$. We used the same LD loss to train both stage 1 and stage 2 model.

In this study, two different types of teacher were utilized depending on the dataset sizes. When the dataset size is sufficiently large, the teacher can be trained sufficiently strong and robust using solely the dataset itself. In this case, the teacher is referred to as the *internal teacher*. Both logit-based and feature-based knowledge distillation methods were used. On the other hand, when the dataset size is limited, training a large-scale, robust teacher model becomes challenging. In such cases, we employed a teacher model trained on large-scale external datasets, referred to as the *external teacher*. While the *external teacher* provides robust feature representations, its predicted logits may not perfectly align with the target dataset's labels due to label variability[6-8,48]. To address this issue, only the feature-based distillation was employed for this case.

The final loss term is calculated as follows:

$$Loss_{Distill} = Loss_{WCE} + w_{LD} Loss_{LD} + w_{FD} Loss_{FD}$$

where $w_{LD} = 2$ and $w_{FD} = 1$ are used for internal teacher and $w_{LD} = 0$ and $w_{FD} = 1$ are used for external teacher.

2.3 Model Training Details

DistillSleep was implemented using Tensorflow [49] framework. We trained both stage 1 and 2 models of *DistillSleep*-T/S over 30 epochs, respectively. AdamW

optimizer was used for training with the learning rate decreasing from 1e-3 to 1e-6 following a cosine annealing schedule. To better account for the inherent uncertainty of the ground truth labels, we used label smoothing [50] during standalone training. Specifically, instead of one-hot encoding, the ground truth probability of class k was defined as following:

$$y_k^{LS} = y_k(1-\alpha) + \alpha/K$$

where y_k is the one-hot encoded value for class k, α is the smoothing parameter and K is total number of classes. We did not use label smoothing during knowledge distillation, since probability vectors from teacher model provide soft target and the effect of label smoothing is limited, as presented in the previous research [51]. To increase data diversity, we used random shifting as data augmentation technique.

2.4 Visualizations of Interpretability

For *intra-epoch* interpretability, we employed Grad-CAM [52] to analyze the contributions of each EEG wavelength range. Grad-CAM calculate the importance of each feature value using gradient information. Specifically, the weight(α^c) for the feature vector of channel c is computed as:

$$\alpha^c = \frac{1}{L}\sum_{i=1}^{L}\frac{\partial \hat{y}_k}{\partial F_i^c}$$

where y_k is the predicted score for the ground-truth class k, F_i^c is the activation of i-th element in the feature vector of channel c and L is the length of the feature vector F^c. The importance of each element $\mathrm{I_i}$ is then calculated as:

$$I_i = Max(0,\ \Sigma_c \alpha^c F_i^c)$$

We applied Grad-CAM to the feature vectors generated by the MWP module.

For *inter-epoch* interpretability, we leveraged the attention scores from the transformer blocks. We examined the global attention scores in the last transformer block of the teacher model, and the attention scores in the single transformer block of the student model. Furthermore, we aggregated attention scores across sliding sequences to investigate broader inter-epoch influences and obtain more robust attention scores.

2.5 Quantization

DistillSleep-S is targeted to be deployed at resource-constrained devices. To further reduce the model size before deployment, *DistillSleep*-S was quantized from real numbers to integer values. Specifically, following Jacob *et al*. [53], we convert all parameters from 32-bit floating point (FLOAT32) to 8-bit integer precision (INT8) using the TensorFlow Lite Converter together with TensorFlow model optimization toolkit [49]. This quantization not only reduces the model's memory requirements but also enables efficient execution on integer-only accelerators, such as mobile NPUs (e.g. Google EdgeTPU). When quantizing the model's weights and activations, two different quantization strategies were considered: Post-Training-Quantization (PTQ) and Quantization-Aware-Training (QAT). PTQ quantizes the weights and activations of the model after the training is completed, based on their distribution. On the other hand, Quantization-Aware-Training (QAT) fine-tunes the model while simulating quantization, adapting the weights for better compatibility with the quantized format. QAT commonly results in lower quantization error and better accuracy, but requires additional effort, as the model must be fine-tuned with quantized parameters.

2.6 Model Calibration

Although deep learning-based models provide fairly high accuracy, they still cannot provide the perfectly precise outcomes. Therefore, it is important to provide the information on how much the model's classification can be trusted so that the clinicians can assess to what extent they can rely on the model's results in clinical practice. The confidence score, defined as the largest value from the model's probability distribution output, can serve as a measure of output reliability. To this end, it must be calibrated to closely reflect the model's true accuracy.

Model calibration [54,55] can be used to evaluate the proximity between the confidence score and model's true accuracy. We assessed the calibration using reliability diagram and expected calibration error (ECE), two of the most commonly used tools for this purpose. In reliability diagram, the expected sample accuracy is plotted as a function of confidence. If the model is perfectly calibrated, which means the confidence perfectly represents the accuracy, the plot exactly follows $y = x$ line. When the model is overconfident, the actual confidence is lower than the accuracy, resulting in the plot positioned below $y = x$ line. Conversely, underconfident models show plots positioned above $y = x$ line. Gap is calculated as the average accuracy of each confidence bin minus $y = x$ line. Thus, positive gap is understood as underconfidence, whereas negative gap is interpreted as overconfidence. ECE represents how much the confidence deviates from accuracy, calculated by the following formula:

$$ECE = \sum_{m=1}^{M}\frac{|B_m|}{n}|acc(B_m) - conf(B_m)|$$

where n is the number of samples, B_m is a confidence bin and m is the number of confidence bin.

3. Results

3.1 Datasets, Preprocessing and Metrics

To train and evaluate *DistillSleep*, we utilized four large-scale datasets that are widely used in sleep research: Sleep Heart Health Study 1 (`SHHS1`) [25,26], Physionet 2018 (`PHY`) [27,28], Danish Center for Sleep Medicine (`DCSM`) and Korea Image Sleep Study (`KISS`) [29].

In addition, we utilized two small datasets to evaluate the

effectiveness of external teacher-based knowledge distillation: SleepEDF-78 [27,33] and Institute of Systems and Robotics, University of Coimbra (ISRUC) [34].

We followed the same train/val/test split used in XSleepNet [22] for SHHS1, SleepEDF-78 and PHY. For DCSM, we followed the train/val/test split used in U-Sleep [20] and for KISS, we followed the train/val/test split used in Jeong *et al.* [29]. For ISRUC, we adhered to the subject group (SG) division specified in the dataset: SG1 (100 subjects, 1 record per subject), SG2 (8 subjects, 2 records per subject), and SG3 (10 subjects, 1 record per subject). For SG1, 50 records were used for knowledge distillation training, 10 records for validation, and the remaining records for testing. SG2 and SG3 were used only for evaluation. The details of each dataset are described in Supplementary Materials.

Following AASM (American Academy of Sleep Medicine) guidelines [56], we used five sleep stages: Wake, N1, N2, N3, and REM. For any datasets scored according to the Rechtschaffen and Kales criteria [57], we rescored the stages by merging N3 and N4 stages into N3 stage. Any epochs scored other than the five sleep stages (e.g. 'MOVEMENT' or 'UNKNOWN') were excluded during preprocessing. Unlike previous studies [22,23,42,45,58] that utilized signal filtering or frequency domain transformation techniques, such as notch filtering, Butterworth filtering, Short-time Fourier Transform, or Fast Fourier Transform, we did not apply these non-trivial methods. Instead, we only standardized the signals and resample them to 200 Hz if the original sampling frequency differs. This not only simplifies the training pipeline but also reduces inference latency by removing the need for complex transformations. For evaluation and comparison to other methods, we used unweighted Macro-F1 score on test split as the major metric. Evaluation results on other metrics (Accuracy, Cohen's kappa, Average sensitivity, Average specificity) are also provided for complementary purposes.

3.2 Teacher Model

3.2.1 Classification Performance

Table 1 compares the performance of *DistillSleep*-T on the four large datasets - PHY, SHHS1, DCSM and KISS - against state-of-the-art methods reported in prior studies [19-24]. *DistillSleep*-T achieved the highest Macro-F1 score across all the datasets, highlighting its effectiveness. While stage 1 model alone delivers competitive predictive performance, the stage 2 model further improves the Macro-F1 score by 3.3-7.1%p. These results demonstrate the ability of the MWP module to capture the characteristics of different wavelengths from the input signal at *intra-epoch* level and that of the hierarchical transformer to aggregate *inter-epoch* level information.

3.2.2 Intra-epoch Interpretability

Figure 2a visualizes the relative importance of different EEG components using Grad-CAM [52] on features after MWP module, based on one-epoch samples for each sleep stage. For Wake and N1 stages, features from smaller pools (Supra-Alpha or Theta) are treated as more significant. In contrast, deeper sleep stages, such as N2 and N3, rely more on features from larger pools (Theta or Delta). This visualization provides clinicians with valuable insights into the mechanisms driving the model's predictions.

To evaluate the consistency of the model's interpretability, we averaged Grad-CAM results from the MWP module across epochs within the same sleep stage (Figure 2b). As sleep deepens from Wake to N1, N2, and N3, the significance of features from larger pools increases. Interestingly, however, in the Wake stage, not only are Supra-Alpha features significant but Infra-Delta features are also highly significant. This finding does not align with the conventional understanding that the Wake stage is dominated by short wavelengths [36-38]. This discrepancy suggests two possibilities: either the Wake-stage data contains meaningful long-wavelength components, or the Infra-Delta max-pooling layer is inadvertently capturing short-wavelength features. To investigate further, we applied a high-pass filter at 0.5 Hz to remove long-wavelength components and re-visualized the Grad-CAM results (Figure 3a). The visualization shows that the significance of Infra-Delta features substantially decreased. Conversely, when the input was low-pass filtered to maintain only long-wavelength components, Infra-Delta features remained strongly activated (Figure 3b). These findings confirm that the MWP module functions as intended, effectively capturing short-wavelength components through small pools and long-wavelength components through large pools. In addition, the observed importance of Infra-Delta features in the Wake stage (Figure 2b) is due to the actual presence of meaningful information in the long-wavelength signals.

To further confirm this observation, we present cases where the model correctly classifies the Wake stage even after applying a low-pass filter (Figure 3c and Figure 3d). Contrary to common assumptions, these cases - supported by both human perception of the raw signals and Grad-CAM results for the MWP module - demonstrate that some Wake-stage epochs are dominated by long wavelengths. This frequency range (0-0.5Hz) has traditionally been filtered out in sleep staging research [18,22,23,42], as it was assumed to contain only noise. However, our analysis reveals the deep learning model's ability to extract informative features from this band. This finding highlights how interpretable models can reveal previously unexplored characteristics inherent in the data, offering new insights into the underlying signals. We will delve deeper into this point in Section 4.

3.2.3 Inter-epoch Interpretability

We used global attention scores from the last transformer block of *DistillSleep*-T to visualize the *inter-epoch* relationships leveraged by the model. Figure 4 provides

attention score visualizations for sequences centered around a sample target epoch. In Figure 4a, the target epoch was initially classified as N1 by the stage 1 model but adjusted to REM by the stage 2 model after accounting for relationships with adjacent epochs. The aggregated attention scores show which neighboring epochs mainly influenced the decision for the target epoch: the 825th and 836th epochs, both classified as REM, consistently exhibit high attention scores, both within individual sliding sequences and in the aggregated attention results. This indicates that these REM epochs contained critical features relevant to REM, providing essential contextual information at the inter-epoch level. On the other hand, when the target epoch was scored as N3 in Figure 4b, no specific neighboring epochs showed strong attention. Instead, the model displays prominent self-attention, implying that the target epoch itself contained distinctive features of N3.

3.3 Student Model

Since the model's representation power typically decreases with smaller size or reduced computational burden as shown by the trendlines in Figure 5b and Figure 5c, it poses a significant challenge to obtain a lightweight model with good predictive performance. As a practical solution for this dilemma, we trained *DistillSleep*-S via knowledge distillation from *DistillSleep*-T. The following subsections present our experimental results; *DistillSleep*-S's feasibility on resource-constrained devices, its predictive performance improvements, and its alignment of internal decision-making process with *DistillSleep*-T through knowledge distillation.

3.3.1 On-device Feasibility

The student model, *DistillSleep*-S, is extremely lightweight having 59.2 times fewer parameters and requiring 19.6 times less computaiton burden compared to *DistillSleep*-T. To validate the practicality of its on-device deployment, we tested *DistillSleep*-S's end-to-end (e2e) latency and memory footprint on resource-constrained devices.

Table 2 presents the e2e latency and peak memory of the INT8-quantized *DistillSleep*-S model for single-epoch classification on various resource-constrained devices (Raspberry Pi 4B, NVIDIA Jetson Orin Nano and Google Coral dev board) as well as a high-performance workstation. Detailed device specifications and measurement process are provided in the Supplementary Materials. For comparison, we also evaluated *DistillSleep*-T. The e2e latency of *DistillSleep*-T was 1473.9 ms on a Raspberry Pi 4B, 569.8 ms on a Jetson Orin Nano, and 210.5 ms even on a high-performance workstation. Additionally, the peak memory consumption reached up to 1191 MB across the tested devices. This makes *DistillSleep*-T, like conventional methods, unsuitable for real-time applications. In contrast, DistillSleep-S demonstrated remarkable efficiency with an e2e latency of 9.1 ms on Raspberry Pi 4B, making it 162 times faster than *DistillSleep*-T. The e2e latency was further reduced when mobile accelerators were used: 5.5 ms on Jetson Orin Nano (mobile GPU) and 8.6 ms on Coral dev board (mobile NPU). In addition, the peak memory remained below 59 MB on tested devices, 20 times lower than *DistillSleep*-T, making *DistillSleep*-S well-suited for edge-devices with limited memory capacity.

3.3.2 Classification Performance

While the student model significantly reduces the model size and computational requirements, this lightweight architecture can lead to losing representational power without a tailored training mechanism. When trained independently, referred to as Baseline-S, the student model exhibited 3.1-5.9%p lower Macro-F1 scores on PHY, SHHS1, DCSM, and KISS compared to *DistillSleep*-T, as shown in Table 3 and Figure 5a. On the other hand, when knowledge distillation was employed, *DistillSleep*-S showed improved classification performance as shown in Table 3. The performance improvement across the four evaluation datasets was 3.4%p on `DCSM`, 2.3%p on `SHHS1`, 2.1%p on `PHY` and 1.9%p on `KISS`.

On smaller datasets `SleepEDF-78` [33] and ISRUC [34] we used *external teacher* for knowledge distillation. This teacher model was trained using 10,723 PSG records sourced from the four large datasets (`PHY`, `SHHS1`, `DCSM`, and `KISS`). For `ISRUC`, `ISRUC-SG1` was used for knowledge distillation training and evaluation, but `ISRUC-SG2` and `ISRUC-SG3` were entirely used for evaluation. Table 3 and Figure 5a demonstrate that *DistillSleep*-S improved the Macro-F1 scores by 5.0%p (`SleepEDF-78`), 4.6%p (`ISRUC-SG1`), 6.8%p (`ISRUC-SG2`), and 2.1%p (`ISRUC-SG3`), thanks to *external teacher*. The greater performance improvement on smaller datasets confirm that knowledge distillation using an external teacher effectively overcomes data limitations to train a robust model.

A class-wise analysis in Table 3 reveals that N1 and REM stages benefit the most from knowledge distillation under both internal and external teachers. This is because, compared to other sleep stages, N1 and REM stages require more comprehensive understanding of both *intra-epoch* and *inter-epoch* information for accurate classification, making them inherently more difficult to classify. The results verify that complex knowledge was successfully transferred from the teacher to the student.

Figure 5b and Figure 5c illustrate the Macro-F1 scores of *DistillSleep*-T and S compared to those of previous methods, plotted against the number of parameters and FLOPs (FLoating point OPerations), respectively. While Baseline-S and the state-of-the art models generally align with the regressed line (dotted line) that is fitted to the number of parameters (or FLOPs) and Macro-F1 scores, *DistillSleep*-S stands significantly above this line, exhibiting superior computational efficiency and tackling the well-known trade-off between model size and accuracy.

Table 4 summarizes *DistillSleep*-S's Macro-F1 scores under different INT8 quantization strategies, on the PHY dataset. Relatively simple Post-Training-Quantization (PTQ) approach resulted in a drastic drop in the Macro-F1 score. While QAT alone improved the Macro-F1 score to 76.6%, combining knowledge distillation during the QAT procedure further boosted the performance to 77.7%, which is only 0.2%p lower than the full-precision (FLOAT32) model.

3.3.3 Interpretability

Since *DistillSleep*-T and *DistillSleep*-S share similar architecture that offers interpretability, we investigated whether the student model inherits internal decision processes of the teacher model through knowledge distillation. To explore this, we compared the interpretability of *DistillSleep*-S and Baseline-S in both the stage 1 and 2 models, using Grad-CAM and attention mechanisms, respectively. Figure 6 shows the averaged Grad-CAM results for *DistillSleep*-S and Baseline-S, providing visualizations analogous to the *DistillSleep*-T results shown in Figure 2b. When comparing the two models, we observed that *DistillSleep*-S exhibits patterns more closely aligned with those of *DistillSleep*-T: for example, *DistillSleep*-S' Theta max-pooled features are more prominent in the N2 stage, and Delta max-pooled features are more highlighted in the N3 stage, compared to Baseline-S. This suggests that knowledge distillation helps align the internal decision processes of *DistillSleep*-S's stage 1 model more closely with those of *DistillSleep*-T.

For stage 2, we examined the attention scores of *DistillSleep*-S and Baseline-S, for the same cases presented in Figure 4. Notably, for the target epoch scored as REM in Figure 7b, Baseline-S shows an attention pattern different from *DistillSleep*-T, requiring greater attention to adjacent epochs. In contrast, in Figure 7a, *DistillSleep*-S shows stronger attention to the same epochs that *DistillSleep*-T focuses on (825th and 836th epochs). For the target epoch scored as N3, the attention patterns in *DistillSleep*-T/S and Baseline-S attention are similar, with relatively weak attention on adjacent epochs. This observation helps explain why N3 benefits relatively less from knowledge distillation, as shown in Table 3.

3.4 Model Robustness and Reliability

3.4.1 Model Generalizability

To assess the model's generalizability to external datasets, we evaluated the external teacher on the SleepEDF-78 and ISRUC datasets, without retraining, as presented in Table 5. In this zero-shot setting, the Macro-F1 score ranged from 55.9 to 67.8%. This performance drop was then mitigated by test-time adaptation, which adjusts the pretrained model using unlabeled test-time data. We adopted two popular methods: Test-time batch normalization (BN) adaptation [59] and Test time ENTropy minimization (TENT) [60]. As shown in Table 5, these methods improved the performance to 66.2-75.8% after test-time BN adaptation and 66.9-76.3% after TENT, in Macro-F1 score. For comparison, a fully finetuned model achieved Macro-F1 of 73.0-81.1%, leaving a final performance gap of 1.7-9.0%p between the adapted model and this upper-bound benchmark.

3.4.2 Robustness to EEG-lead choice

To investigate the robustness of *DistillSleep* to the choice of EEG lead, we evaluated its performance on the PHY dataset using two additional EEG leads (F3-M2 and O1-M2) and compared the results to the C4-M1 lead used in our main study. The experimental results are summarized in Table 6. Consistent with previous studies [18,20], the choice of EEG lead did not result in significant variations in predictive performance. For *DistillSleep*-T, compared to the C4-M1 lead, the F3-M2 lead resulted in 0.6%p increase in Macro-F1 score, while O1-M2 lead resulted in 1.9%p drop. The suboptimal performance from an occipital lead is also consistent with prior research. A similar tendency was observed for *DistillSleep*-S and Baseline-S, where the F3-M2 lead exhibited slightly improved predictive performance and the O1-M2 lead showed lower predictive performance.

3.4.3 Model Reliability

Figure 8 exhibits the reliability diagrams of *DistillSleep*-T, *DistillSleep*-S, and Baseline-S, along with the corresponding ECE, measured on PHY. In these diagrams, *DistillSleep*-T shows a strong alignment between confidence scores and accuracy, although its confidence levels are slightly lower than the actual accuracy. Baseline-S exhibits a calibration gap of a similar magnitude to *DistillSleep*-T. In contrast, *DistillSleep*-S achieves near-perfect calibration: the gap in the reliability diagram is significantly reduced, and its ECE is 2.7$\mathbb{X}$ smaller than that of both *DistillSleep*-T and Baseline-S. These findings suggest that knowledge distillation improves the calibration of the student model by combining the richer information from both training data and the teacher model.

4. Discussion

Over the past decade, deep-learning systems such as DeepSleepNet [18], IITNet [19], U-Time [21]/U-Sleep [20], XSleepNet [22], SleepTransformer [23] and, most recently, SleePyCo [24] have steadily raised the ceiling on single-channel sleep-staging accuracy. However, these models remain computationally heavy. For example, SleePyCo requires 2.3 million parameters and 11.2 GFLOPs for inference. Such footprints preclude real-time inference on bedside monitors and wearable.

Since the model's representation power typically increases with greater size or heavier computational burden as shown in Figure 5b and Figure 5c, it poses a significant challenge to obtain the lightweight model with good predictive performance. As a practical solution for

this dilemma, *DistillSleep* introduces a dual-model framework consisting of performance-oriented teacher and lightweight student. The teacher model outperforms other existing models in terms of predictive performance. This model can be used where the accurate prediction is important, such as automatic PSG scoring. The student model, on the other hand, is very lightweight to complete the single epoch inference in 9.1ms on Raspberry Pi 4B. Despite its small size, the predictive performance of *DistillSleep*-S remains competitive thanks to knowledge distillation from the strong teacher model. Moreover, the external teacher, trained on 10,723 PSG records, is readily employed to train a robust student model even with as few as 50 training samples. This approach eliminates the challenges of training a strong teacher model with limited data, enhancing the practicality. The student model can be readily used for broader applications where computation resources are limited or having timely results is important. Such cases include breathing control for patients with obstructive sleep apnea (OSA) or patients on mechanical ventilators [11] [12], remote sleep monitoring for optimized care delivery [13], temperature control during sleep [14,15], automatically adjustable bed [16], and memory consolidation [17].

Moreover, *DistillSleep* provides interpretability at both *intra-* and *inter-epoch* levels through the MWP module and attention mechanisms. These features allow human users to trace the model's internal decision making processes, making *DistillSleep* more reliable and transparent. For example, according to AASM rules, Delta waves should constitute at least 20% of an epoch to be scored as N3 stage. By observing MWP's Delta activation in N3-scored epochs, clinicians can verify if the model is focusing on Delta waves as expected. Another example includes verifying REM stages using attention scores. AASM rules state that if any prior epoch is scored as REM and no stopping criteria exist in the following epochs, the target epoch should be scored as REM even in the absence of clear REM indicators in it. By examining attention scores as in Figure 4a, clinicians can confirm whether these rules are consistently applied.

Additionally, interpretability analysis offers deeper insight into the unexplored potential of deep learning models. As demonstrated in Section 3.2.2, the MWP module provides an understanding of how the model utilizes slow waves in the 0-0.5 Hz band - previously dismissed as irrelevant to sleep staging - for classifying the Wake stage. This counterintuitive observation may stem from the model's reliance on single-channel EEG data, unlike the AASM guidelines, which utilize multimodal full PSG signals. Even in scenarios where other signals, such as EEG from different positions, EOG, EMG, or ECG play a more critical role than the target EEG channel in conventional sleep staging, single-channel models are forced to extract all relevant features solely from the target EEG channel. Consequently, transparent interpretation of these models can illuminate how information from other modalities is *indirectly encoded* within the target EEG signals and which EEG features are prioritized when other channels are eliminated, offering unique insights on the model's decision-making process.

Previous research examined the EEG components in infra-Delta band, known as infra-slow oscillations (ISOs) [61] [aladjalova1957infra]. Studies have shown that these oscillations are related to autonomic control system [62] [63] and are sensitive to stress from cognitive tasks [64] [prokhorov2023changes]. Our results indicate that deep learning models may differentiate patterns of ISOs between waking and sleeping states. Another possible explanation is that this Wake-stage classification may come from combined eye movements or blinking, as stated in the AASM guidelines [56].

Although these are not explicit EEG scoring criteria, *DistillSleep* may effectively identify and utilize inherent EEG features related to these activities from the raw data for its Wake-stage classification. The above example demonstrates the value of an interpretable model. The ‘black-box’ nature of deep learning models veils which data characteristics serve as meaningful predictive features, resulting in the exclusion of potentially important information from the raw data by following conventional preprocessing. For instance, filtering out 0-0.5Hz frequency range from the raw signal, as done in prior studies [18,22,23,42], results in misclassification of certain samples as presented in Figure 3c and Figure 3d. In contrast, interpretable models enable users to not only gain insights into the underlying mechanisms of the model's predictions but also identify the important data characteristics, ultimately enabling more effective preprocessing strategies.

Previous studies have proposed interpretable sleep staging models, employing techniques such as layerwise relevance propagation on time-frequency images [65], attention scores in transformer blocks [23] [66] or eigen-CAM on screen-captured PSG images [29] [43]. However, these methods focused on large-scale models and did not explore whether the interpretability is maintained after model compression. Our interpretability analysis verifies that when trained independently, a small model focuses on different features compared to the large-scale teacher model. In contrast, *DistillSleep*-S addresses the limitation via knowledge distillation, effectively preserving the internal decision-making processes of the teacher model after model size reduction, as shown in Figure 6 and Figure 7. These findings indicate that knowledge distillation not only enhances *DistillSleep*-S's predictive performance but also ensures consistent interpretative patterns with *DistillSleep*-T for identical inputs. This synchronized behavior allows practitioners to confidently deploy either model interchangeably based on specific task requirements, knowing the underlying reasoning remains consistent. For example, *DistillSleep*-S can be utilized for latency-sensitive applications while *DistillSleep*-T can be employed for scenarios demanding higher accuracy.

In addition, an improvement in model calibration was

observed in *DistillSleep*-S, addressing the underconfidence of its teacher. Sleep staging inherently involves a high degree of uncertainty, making it crucial to effectively capture this uncertainty for both better performance and calibration [67,68]. Knowledge distillation, with its use of soft labels, naturally embeds uncertainty, making it well-suited for this inherently uncertain task. Interestingly, while previous works have noted that knowledge distillation mitigates overconfidence in student models [51], our results show that it can also improve underconfidence. To our knowledge, this is the first study to apply knowledge distillation for sleep staging, while demonstrating that interpretability is effectively transferred and model calibration is improved. Even when stronger models can emerge in the future, whether in terms of predictive performance, interpretability or model calibration, our approach may expand their applicability across varying objectives via knowledge distillation.

DistillSleep equips clinicians with practical verification tools, including intra-epoch wavelength analysis, inter-epoch attention scores, and confidence scores to validate and interpret the scoring results. In practice, confidence scores can serve as an initial filter, identifying epochs with low confidence scores for further manual review. Following this initial filtering, interpretability tools designed for both *intra-epoch* and *inter-epoch* analysis can be flexibly applied depending on the specific scenario. When both the stage 1 and 2 models provide consistent sleep staging results, intra-epoch features suffice to explain the model's prediction. In such cases, the MWP module offers insights into the contributions of different wavelength components to the decision-making process, as detailed in Section 3.2.2. Conversely, if the stage 2 model's predictions differ from those of the stage 1 model, inter-epoch interpretability becomes essential. In such cases, attention scores help identify which adjacent epochs significantly influence the stage 2 model's final prediction. The combination of verification tools makes *DistillSleep* a robust and transparent framework, ultimately enhancing trust and facilitating the integration of automatic sleep staging into clinical practice.

Modeling inter-epoch relations enhances prediction accuracy as presented in Table 1, but achieving immediate results in a real-time deployment requires special consideration. While the stage 2 model of DistillSleep-S aggregates information across multi-epochs, we configure it to rely only on the current and preceding epochs. This allows each new epoch to be classified immediately, without waiting for future data. Furthermore, the hybrid design of *DistillSleep*-S - combining a one-to-one framework in stage 1 with a many-to-many framework in stage 2 - is better suited for deployment than the single model design used by most state-of-the-art methods [20-23]. Single model architecture poses two main limitations in practical applications: start-up delays and computational inefficiency. First, single many-to-many models must wait to accumulate a full sequence before producing predictions, leading to delays at the beginning of the sleep monitoring or after interruptions like sensor detachment. In contrast, the one-to-one stage 1 model in *DistillSleep*-S provides immediate, per-epoch classification results, eliminating start-up delays. Second, single-model designs recompute the entire pipeline for every new sequence, even if the majority of input data overlaps. This redundancy incurs substantial computational inefficiency. Conversely, *DistillSleep*-S avoids this inefficiency through stage separation: the computationally intensive intra-epoch feature extraction (stage 1) is performed only once per epoch, and the resulting features are cached and reused by the stage 2 model as the input sequence slides forward. This caching mechanism eliminates redundant computation on overlapping data and, as measured in our experiments, reduces the e2e latency eightfold from 70.9 ms to 9.1 ms on a Raspberry Pi 4B. With these deployment-oriented considerations, *DistillSleep*-S achieves real-time e2e latency, as presented in Table 2.

Our robustness tests on unseen data variability demonstrate that, when combined with test-time adaptation strategies, *DistillSleep* can maintain competitive predictive performance despite distributional shifts in the data. We attribute a large part of the remaining performance gap between the adapted model and the fully fine-tuned model to label variability, which cannot be directly corrected by these adaptation methods. Notably, the gap is larger for `SleepEDF-78`, likely due to its use of older Rechtschaffen and Kales criteria [57], which can lead to greater label discrepancies. A limitation of our current study is that we do not explicitly address this label variability.

Consistent with previous works [18] [20], our analysis showed that the choice of EEG lead did not result in significant variations in the performance of *DistillSleep*. However, a limitation of our study is that we did not perform a deeper analysis to understand the precise reasons for these variations.

While previous studies have included the datasets collected from restricted regions (Europe or North America), our study expands the regional diversity to East Asia, by including the `KISS` dataset. By doing so, we validate that *DistillSleep* can work well regardless of regional demographic biases. Still, there remains regions not covered, and we hope more data are collected from other regions and publicly shared. The datasets used in our study include only healthy individuals and patients diagnosed with OSA. However, any subjects exhibiting abnormal brain activity are not included, such as those with a history of stroke, psychiatric conditions, or neurodegenerative disorders. Broader study is necessary to confirm whether *DistillSleep* still could work well on the datasets collected from such subjects.

Acknowledgements

This research was supported partly by Creative-

Pioneering Researchers Program through Seoul National University, partly by AI-Bio Research Grant through Seoul National University, partly by a grant of `Korea Government Grant Program for Education and Research in Medical AI' through the Korea Health Industry Development Institute (KHIDI), funded by the Korea government (MOE, MOHW), partly by the National research Foundation (NRF) of Korea grant funded by the Korea government (MSIT) (No. RS-2023-00222663). This research was also supported by the data construction project for artificial intelligence through the National Information Agency of Korea (NIA) funded by the Ministry of Science and ICT.

Disclosure statements

Financial disclosure: H.-W.S. is an inventor on patent applications submitted by Seoul National University related to an image-based polysomnography dataset and its application. H.-W.S. is a founder of OUaR LaB, Inc., serves on the Board of Directors and as a chief executive officer for OUaR LaB, Inc., and owns OUaR LaB Stock, which are subject to certain restrictions under university policy. All other authros declares no competing interests. Nonfinancial disclosure: Authors have no conflict of nonfinancial interest to declare.

References

1. Benjafield AV, Ayas NT, Eastwood PR, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *The Lancet respiratory medicine*. 2019;7(8):687–698.

2. Peppard PE, Young T, Barnet JH, Palta M, Hagen EW, Hla KM. Increased prevalence of sleep-disordered breathing in adults. *American journal of epidemiology*. 2013;177(9):1006–1014.

3. Hafner M, Stepanek M, Taylor J, Troxel WM, Van Stolk C. Why sleep matters—the economic costs of insufficient sleep: a cross-country comparative analysis. *Rand health quarterly*. 2017;6(4):11.

4. Ryu S, Kim SC, Kim RB, et al. Influence of Sleep Stage on the Determination of Positional Dependency in Patients With Obstructive Sleep Apnea. *Clinical and Experimental Otorhinolaryngology*. 2024;17(3):226–233.

5. Sateia MJ. International classification of sleep disorders. *Chest*. 2014;146(5):1387–1394.

6. Danker-Hopfe H, Kunz D, Gruber G, et al. Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders. *Journal of sleep research*. 2004;13(1):63–69.

7. Danker-Hopfe H, Anderer P, Zeitlhofer J, et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *Journal of sleep research*. 2009;18(1):74–84.

8. Guillot A, Sauvet F, During EH, Thorey V. Dreem open datasets: Multi-scored sleep datasets to compare human and automated sleep staging. *IEEE transactions on neural systems and rehabilitation engineering*. 2020;28(9):1955–1965.

9. Gottlieb DJ, Punjabi NM. Diagnosis and management of obstructive sleep apnea: a review. *Jama*. 2020;323(14):1389–1400.

10. Huttunen R, Leppänen T, Duce B, et al. Assessment of obstructive sleep apnea-related sleep fragmentation utilizing deep learning-based sleep staging from photoplethysmography. *Sleep*. 2021;44(10):zsab142.

11. Lee E-M, Lee T-H, Park O-L, Nam JG. Effective continuous positive airway pressure changes related to sleep stage and body position in obstructive sleep apnea during upward and downward titration: an experimental study. *Journal of Clinical Neurology*. 2020;16(1):90–95.

12. On S-W, Kim D-K, Lee MH, et al. Clinical Efficacy of a Position-Responding Mandibular Advancement Device in Patients With Obstructive Sleep Apnea. *Clinical and Experimental Otorhinolaryngology*. 2024;17(4):302–309.

13. Tajima Y, Nakata M, Takadama K. Personalized real-time sleep stage remote monitoring system. IEEE; 2014:1–5.

14. Ngarambe J, Yun GY, Lee K, Hwang Y. Effects of changing air temperature at different sleep stages on the subjective evaluation of sleep quality. *Sustainability*. 2019;11(5):1417.

15. Okamoto-Mizuno K, Mizuno K. Effects of thermal environment on sleep and circadian rhythm. *Journal of physiological anthropology*. 2012;31(1):14.

16. Wei Y, Zhu Y, Zhou Y, et al. Investigating the influence of an adjustable zoned air mattress on sleep: a multinight polysomnography study. *Frontiers in Neuroscience*. 2023;17:1160805.

17. Leminen MM, Virkkala J, Saure E, et al. Enhanced memory consolidation via automatic sound stimulation during non-REM sleep. *Sleep*. 2017;40(3):zsx003.

18. Supratak A, Dong H, Wu C, Guo Y. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE transactions on neural systems and rehabilitation engineering*. 2017;25(11):1998–2008.

19. Seo H, Back S, Lee S, Park D, Kim T, Lee K. Intra-and inter-epoch temporal context network (IITNet) using sub-epoch features for automatic sleep scoring on raw single-channel EEG. *Biomedical signal processing and control*. 2020;61:102037.

20. Perslev M, Darkner S, Kempfner L, Nikolic M, Jennum PJ, Igel C. U-Sleep: resilient high-frequency sleep staging. *NPJ digital medicine*. 2021;4(1):72.

21. Perslev M, Jensen M, Darkner S, Jennum PJ,

Igel C. U-time: A fully convolutional network for time series segmentation applied to sleep staging. *Advances in neural information processing systems*. 2019;32

22. Phan H, Chén OY, Tran MC, Koch P, Mertins A, De Vos M. XSleepNet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2021;44(9):5903–5915.

23. Phan H, Mikkelsen K, Chén OY, Koch P, Mertins A, De Vos M. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering*. 2022;69(8):2456–2467.

24. Lee S, Yu Y, Back S, Seo H, Lee K. Sleepyco: Automatic sleep scoring with feature pyramid and contrastive learning. *Expert Systems with Applications*. 2024;240:122551.

25. Zhang G-Q, Cui L, Mueller R, et al. The National Sleep Research Resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*. 2018;25(10):1351–1358.

26. Quan SF, Howard BV, Iber C, et al. The sleep heart health study: design, rationale, and methods. *Sleep*. 1997;20(12):1077–1085.

27. Goldberger AL, Amaral LA, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation*. 2000;101(23):e215–e220.

28. Ghassemi MM, Moody BE, Lehman L-WH, et al. You snooze, you win: the physionet/computing in cardiology challenge 2018. IEEE; 2018:1–4.

29. Jeong J, Yoon W, Lee J-G, et al. Standardized image-based polysomnography database and deep learning algorithm for sleep-stage classification. *Sleep*. 2023;46(12):zsad242.

30. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:150302531*. 2015;

31. Heo B, Kim J, Yun S, Park H, Kwak N, Choi JY. A comprehensive overhaul of feature distillation. 2019:1921–1930.

32. Huang T, You S, Wang F, Qian C, Xu C. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems*. 2022;35:33716–33727.

33. Kemp B, Zwinderman AH, Tuk B, Kamphuisen HA, Oberye JJ. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*. 2000;47(9):1185–1194.

34. Khalighi S, Sousa T, Santos JM, Nunes U. ISRUC-Sleep: A comprehensive public dataset for sleep researchers. *Computer methods and programs in biomedicine*. 2016;124:180–192.

35. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. Mobilenetv2: Inverted residuals and linear bottlenecks. 2018:4510–4520.

36. Armitage R. The distribution of EEG frequencies in REM and NREM sleep stages in healthy young adults. *Sleep*. 1995;18(5):334–341.

37. Huang H, Zhang J, Zhu L, et al. EEG-based sleep staging analysis with functional connectivity. *Sensors*. 2021;21(6):1988.

38. Purves D, Augustine GJ, Fitzpatrick D, Hall W, LaMantia A-S, White L. *Neurosciences*. De Boeck Supérieur; 2019.

39. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Advances in neural information processing systems*. 2017;30

40. Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows. 2021:10012–10022.

41. Liu Z, Hu H, Lin Y, et al. Swin transformer v2: Scaling up capacity and resolution. 2022:12009–12019.

42. Phan H, Andreotti F, Cooray N, Chén OY, De Vos M. SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2019;27(3):400–410.

43. Lee H, Choi YR, Lee HK, et al. Explainable vision transformer for automatic visual sleep staging on multimodal PSG signals. *npj Digital Medicine*. 2025;8(1):55.

44. Eldele E, Chen Z, Liu C, et al. An attention-based deep learning approach for sleep stage classification with single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2021;29:809–818.

45. Goshtasbi N, Boostani R, Sanei S. SleepFCN: A fully convolutional deep learning framework for sleep stage classification using single-channel electroencephalograms. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2022;30:2088–2096.

46. Peng B, Jin X, Liu J, et al. Correlation congruence for knowledge distillation. 2019:5007–5016.

47. Shin S, Lee J, Lee J, Yu Y, Lee K. Teaching where to look: Attention similarity knowledge distillation for low resolution face recognition. Springer; 2022:631–647.

48. Moser D, Anderer P, Gruber G, et al. Sleep classification according to AASM and Rechtschaffen & Kales: effects on sleep scoring parameters. *Sleep*. 2009;32(2):139–149.

49. Abadi M, Agarwal A, Barham P, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:160304467*. 2016;

50. Szegedy C, Vanhoucke V, Ioffe S, Shlens J,

Wojna Z. Rethinking the inception architecture for computer vision. 2016:2818–2826.

51. Müller R, Kornblith S, Hinton GE. When does label smoothing help? *Advances in neural information processing systems*. 2019;32

52. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. 2017:618–626.

53. Jacob B, Kligys S, Chen B, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference. 2018:2704–2713.

54. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. PMLR; 2017:1321–1330.

55. Mehrtash A, Wells WM, Tempany CM, Abolmaesumi P, Kapur T. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE transactions on medical imaging*. 2020;39(12):3868–3878.

56. Berry RB, Brooks R, Gamaldo C, et al. AASM scoring manual updates for 2017 (version 2.4). American Academy of Sleep Medicine; 2017. p. 665–666.

57. Rechtschaffen A. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. *NIH publication (US Department of Health, Education, and Welfare)*. 1968:1–55.

58. Chambon S, Galtier MN, Arnal PJ, Wainrib G, Gramfort A. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2018;26(4):758–769.

59. Schneider S, Rusak E, Eck L, Bringmann O, Brendel W, Bethge M. Improving robustness against common corruptions by covariate shift adaptation. *Advances in neural information processing systems*. 2020;33:11539–11551.

60. Wang D, Shelhamer E, Liu S, Olshausen B, Darrell T. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:200610726*. 2020;

61. Aladjalova N. Infra-slow rhythmic oscillations of the steady potential of the cerebral cortex. *Nature*. 1957;179(4567):957–959.

62. Knyazev GG. EEG delta oscillations as a correlate of basic homeostatic and motivational processes. *Neuroscience & Biobehavioral Reviews*. 2012;36(1):677–695.

63. Karavaev A, Kiselev A, Runnova A, et al. Synchronization of infra-slow oscillations of brain potentials with respiration. *Chaos: An Interdisciplinary Journal of Nonlinear Science*. 2018;28(8)

64. Prokhorov MD, Borovkova EI, Hramkov AN, et al. Changes in the power and coupling of infra-slow oscillations in the signals of EEG leads during stress-inducing cognitive tasks. *Applied Sciences*. 2023;13(14):8390.

65. Zhou D, Xu Q, Zhang J, et al. Interpretable sleep stage classification based on layer-wise relevance propagation. *IEEE Transactions on Instrumentation and Measurement*. 2024;73:1–10.

66. Pradeepkumar J, Anandakumar M, Kugathasan V, et al. Towards interpretable sleep stage classification using cross-modal transformers. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2024;

67. Van Gorp H, van Gilst MM, Fonseca P, Overeem S, van Sloun RJ. Modeling the impact of inter-rater disagreement on sleep statistics using deep generative learning. *IEEE Journal of Biomedical and Health Informatics*. 2023;27(11):5599–5609.

68. Heremans ER, Seedat N, Buyse B, Testelmans D, van der Schaar M, De Vos M. U-PASS: An uncertainty-guided deep learning pipeline for automated sleep staging. *Computers in Biology and Medicine*. 2024;171:108205.

Figures

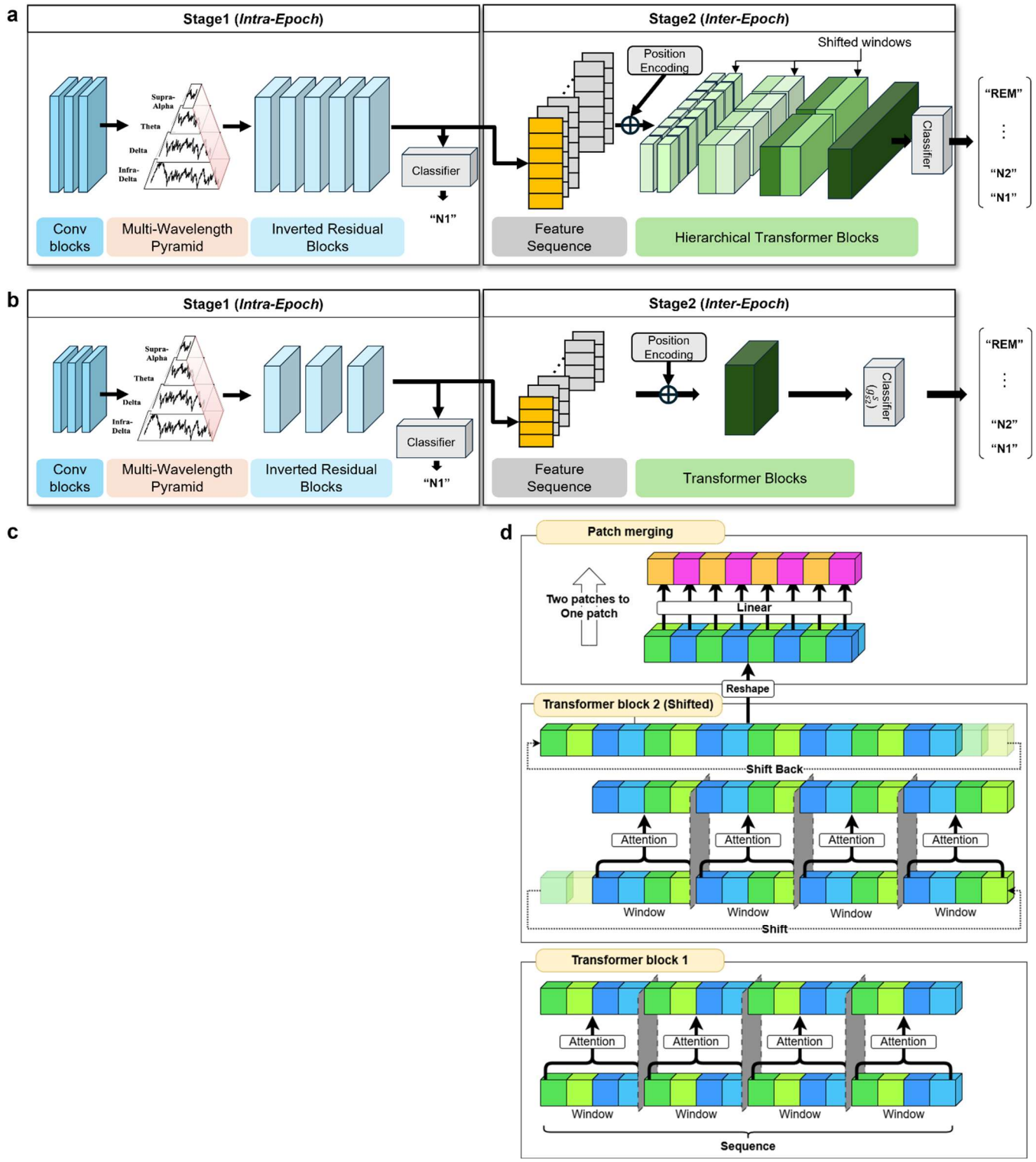

Figure 1. Model architecture. a. Teacher model is structured in two-stage architecture. The first stage captures *intra-epoch* features, while the second stage aggregates *inter-epoch* information. The first stage model consists of convolutional blocks, MWP module and inverted residual blocks. The second stage model is built on hierarchical transformer architecture. **b**. The student model shares a similar architecture with the teacher model, but has fewer layers and filters. The second stage model of the student model has a single transformer block. **c**. Multi-Wavelength Pyramid (MWP) captures features from different range of wavelengths from the input signal by applying multiple max-pooling operations in parallel. The max-poolings are termed as Supra-Alpha, Theta, Delta, and Infra-Delta to reflect their respective EEG types of interest. **d.** Hierarchical transformers in the teacher model efficiently aggregates *inter-epoch* relationships. Each hierarchical level consists of two transformer blocks with local attention and patch merging. The input sequence is shifted between two transformer blocks. Patch merging aggregates adjacent patches for higher-level features.

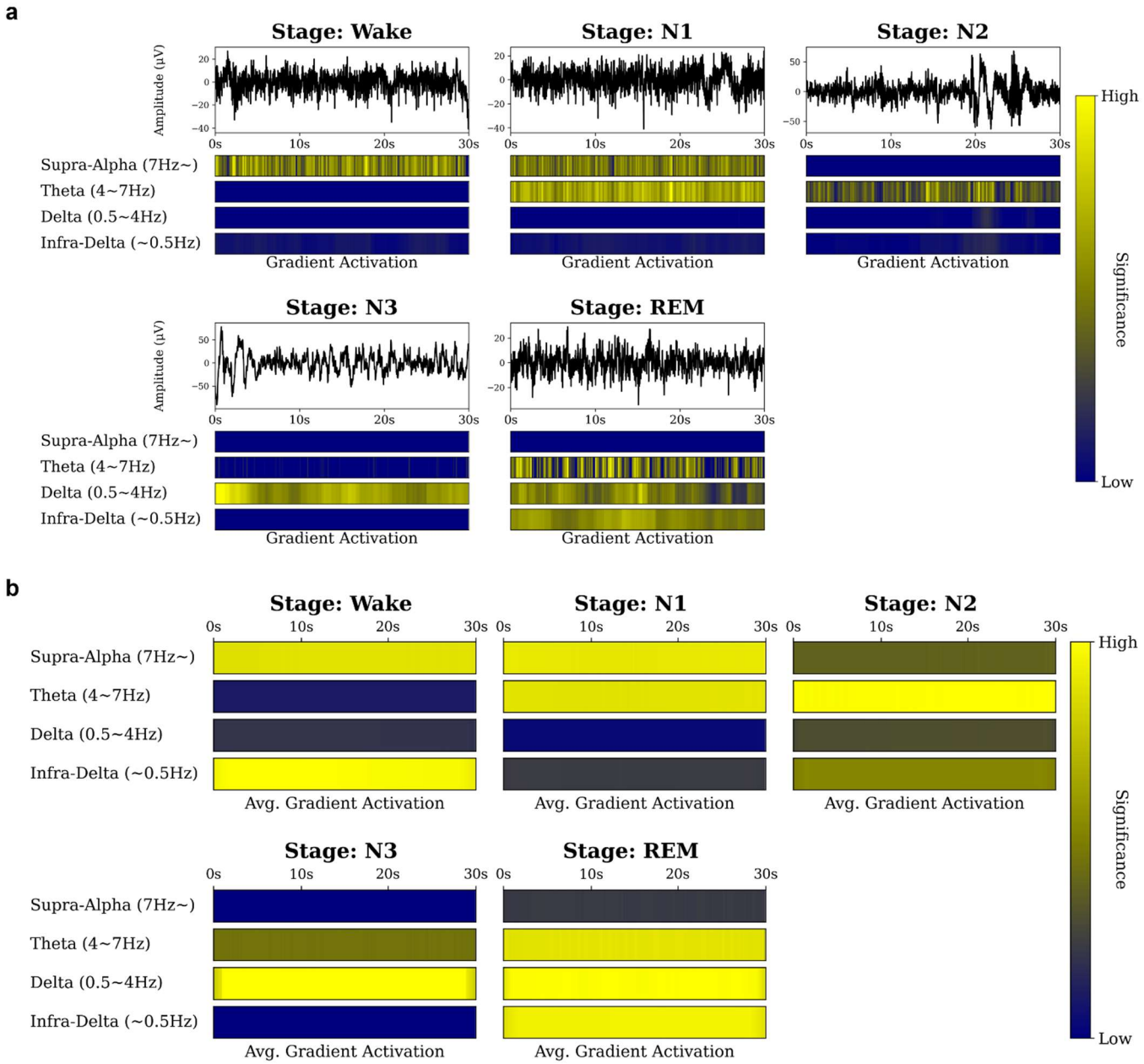

Figure 2. Intra-epoch interpretability. a. Grad-CAM visualization of MWP features on sample signals from each stage. Wavelength components considered significant by the model are highlighted and can be used for verification. In samples from the Wake and N1 stages, features from smaller pools (Supra-Alpha or Theta) are highlighted, whereas for deeper sleep stages, larger pools (Theta or Delta) are emphasized. **b.** Averaged Grad-CAM results per each stage. As sleep deepens, the model tends to place more significance on max-pools with larger pool sizes. The data is from `PHY` dataset. To avoid confusion from misclassified epochs, we only include correctly classified epochs in the averaging.

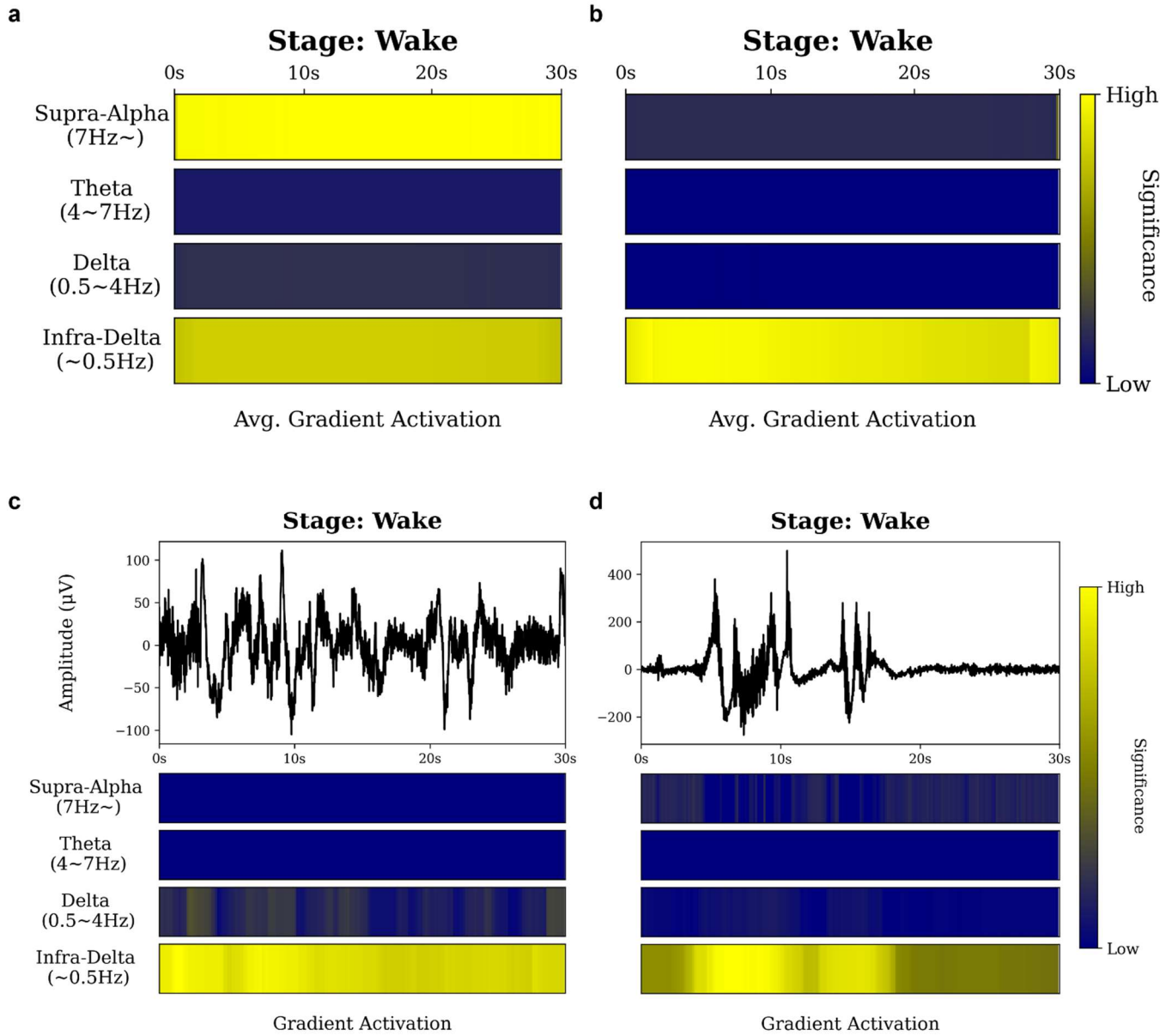

Figure 3. Averaged Grad-CAM results on Wake stage after high/lowpass filter and samples with strong Infra-Delta component. a-b. Removing long/short wavelength components reduces the significance of large/small maxpools, which verifies the *interpretability* of the MWP. FIR filter is used with 0.5Hz as cutoff frequency for both highpass and lowpass filter. **a.** Averaged Grad-CAM Results on highpass-filtered signal. The significance of Infra-Delta max-pooled features substantially decreases. **b.** Averaged Grad-CAM Results on lowpass-filtered signal. Contrary to highpass-filter case, the significance of Supra-Alpha max-pooled features diminishes. **c-d.** These samples demonstrate that some epochs from Wake stage are actually dominated by slow waves, which are used by the model for classification.

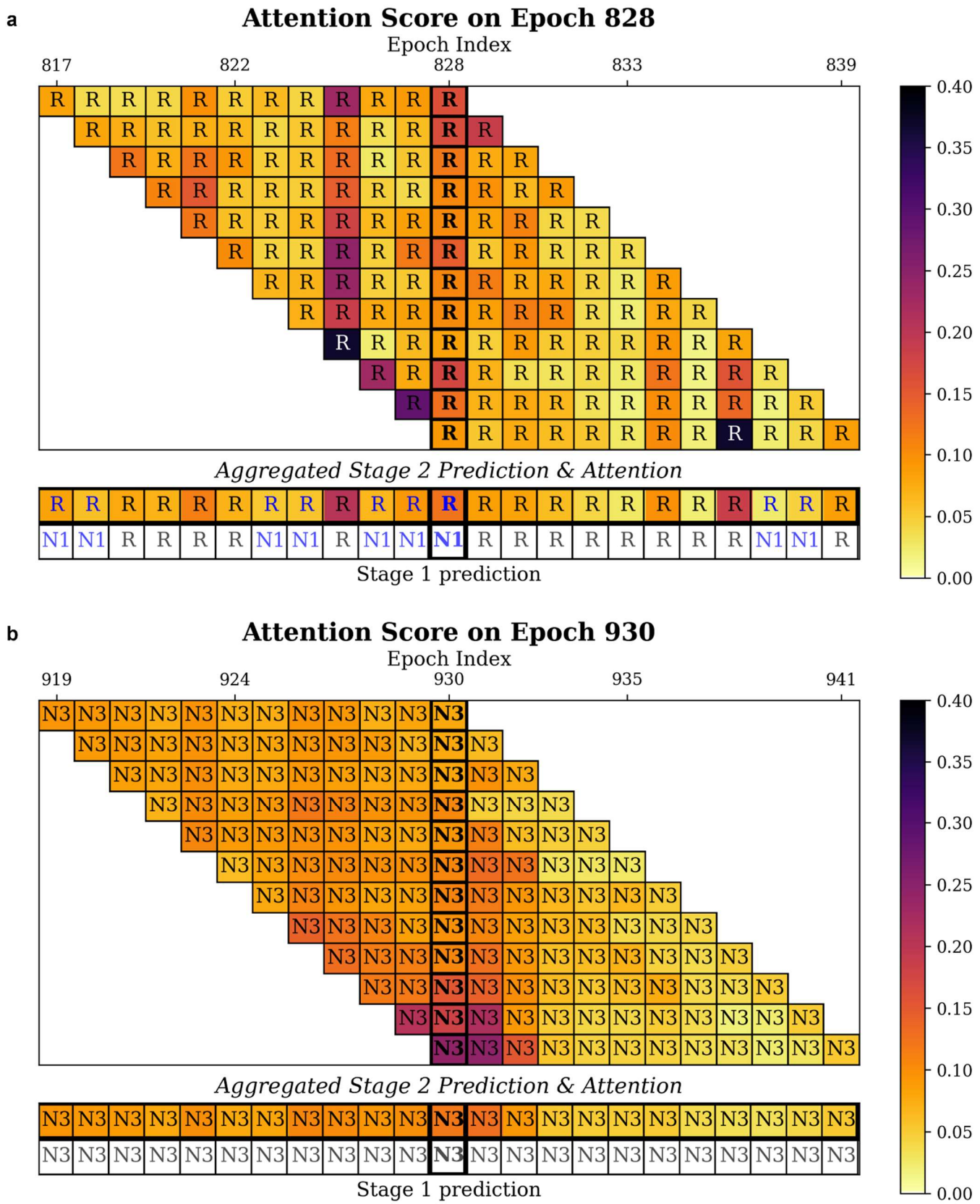

Figure 4. Inter-epoch interpretability using attention scores. Attention scores can be used to understand which specific epoch influences the target epoch's scoring decision. In each subfigure, the top part represents attention scores on sliding sequences and the bottom part shows the aggregated attention scores. Stage 1 prediction is provided under the aggregated attention scores. Blue text indicates epochs where the stage 2 model corrected the stage 1 model's prediction. **a.** Attention score on REM scored sample. **b.** Attention score on N3 scored sample.

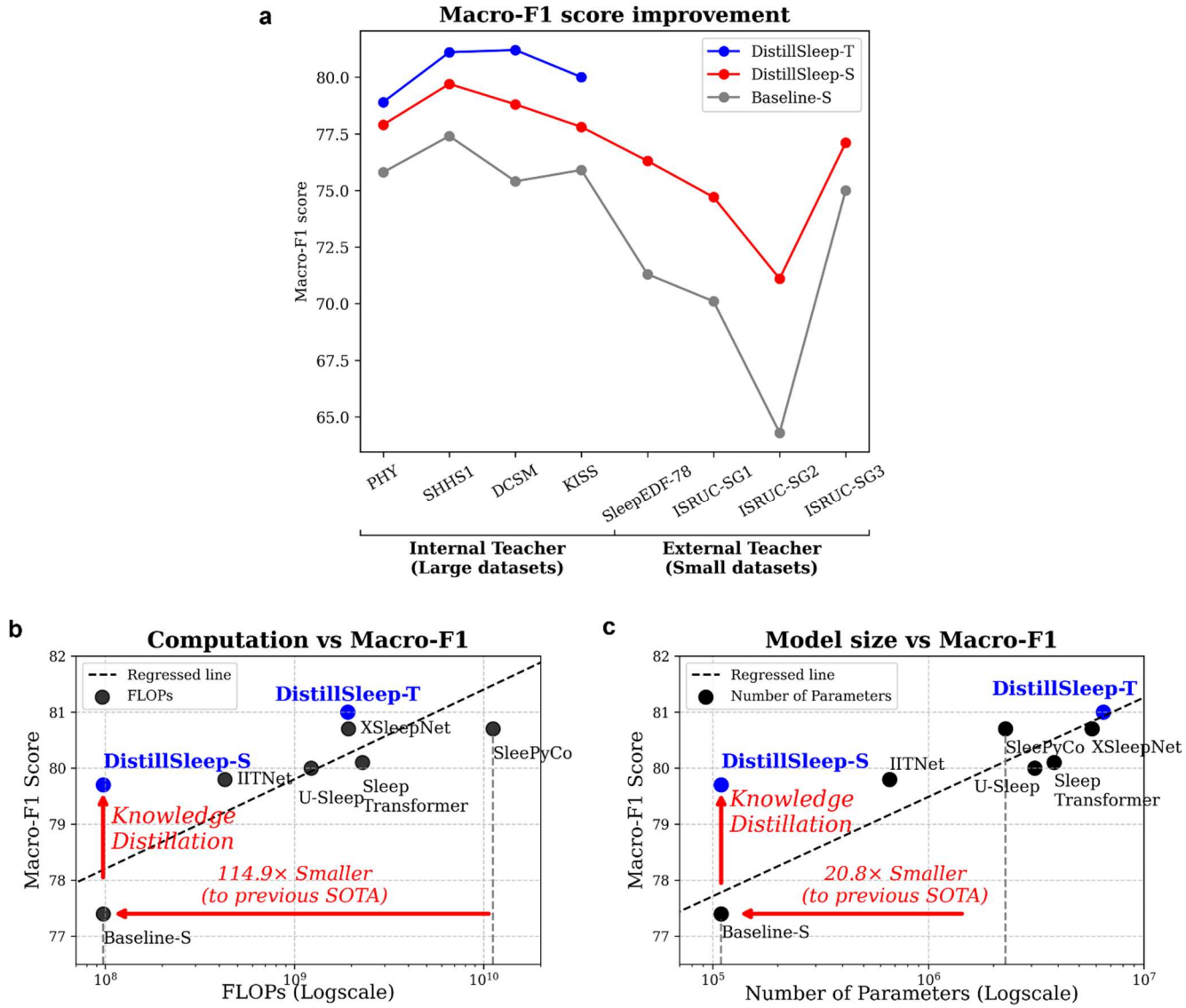

Figure 5. Effect of knowledge distillation. a. Knowledge distillation improves the Macro-F1 score of *DistillSleep*-S by 1.9-3.4%p, when internal teacher is used. Even when the teacher trained on external datasets is used (external teacher), the strong teacher can improve the Macro-F1 score of *DistillSleep*-S up to 6.8%p. **b-c**. *DistillSleep*-S achieves competitive predictive performance despite its smaller size and lower computational cost, thanks to knowledge distillation from strong *DistillSleep*-T. As the figures show, *DistillSleep*-S significantly outperforms the expected Macro-F1 score (dotted line) based on its size and computational budget. This contrasts with both the baseline model (Baseline-S) and state-of-the-art models, including *DistillSleep*-T, which adhere closely to the regression line. Macro-F1 on `SHHS1` is used for plotting.

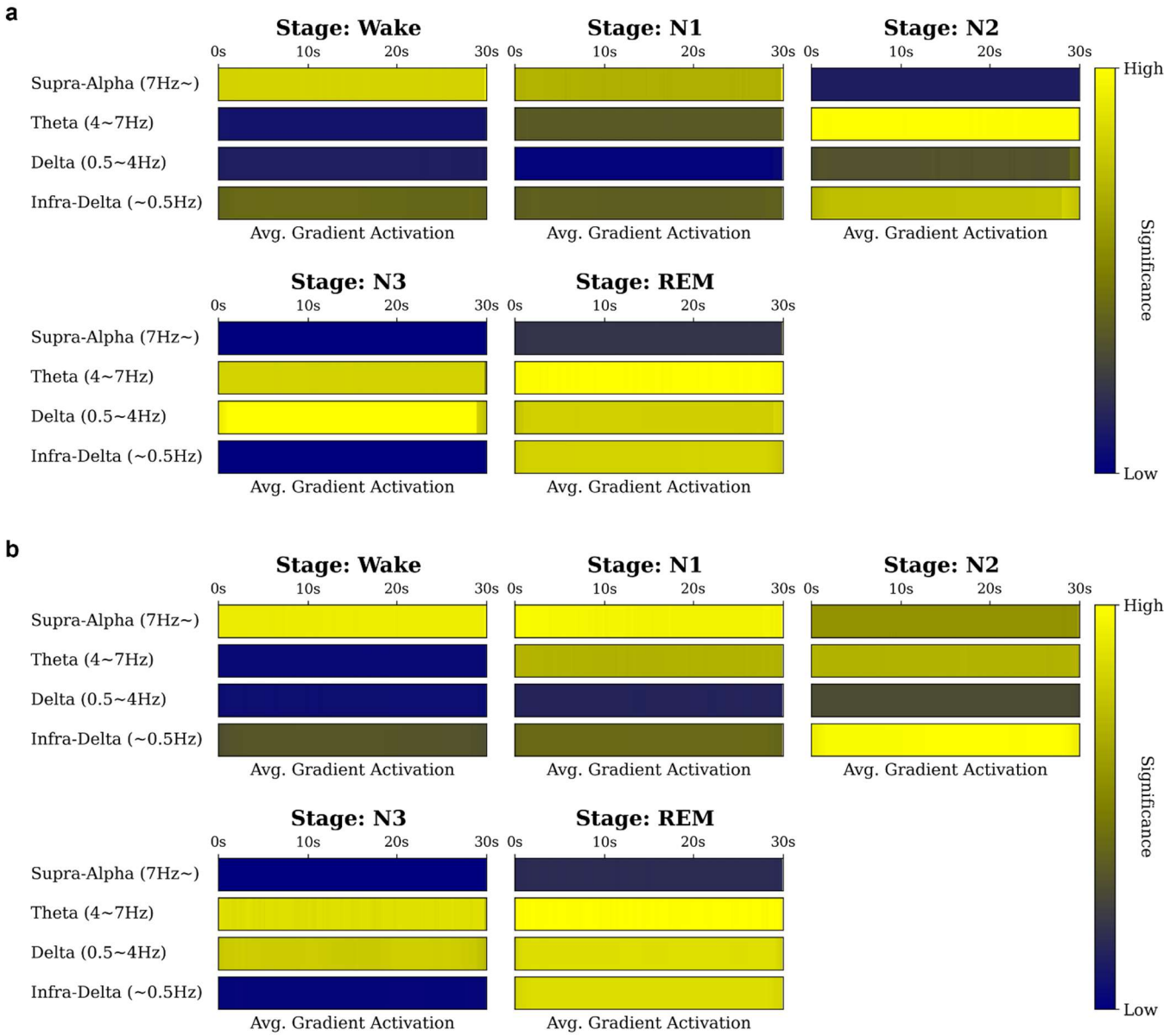

Figure 6. Averaged Grad-CAM results of *DistillSleep*-S and Baseline-S. Knowledge distillation adjusts *DistillSleep*-S to have similar internal decision process to *DistillSleep*-T, as demonstrated from the visualization. The result is based on `PHY` dataset. More results on other datasets are presented in Supplementary materials. **a.** Averaged Grad-CAM results of *DistillSleep*-S. **b.** Averaged Grad-CAM results of Baseline-S.

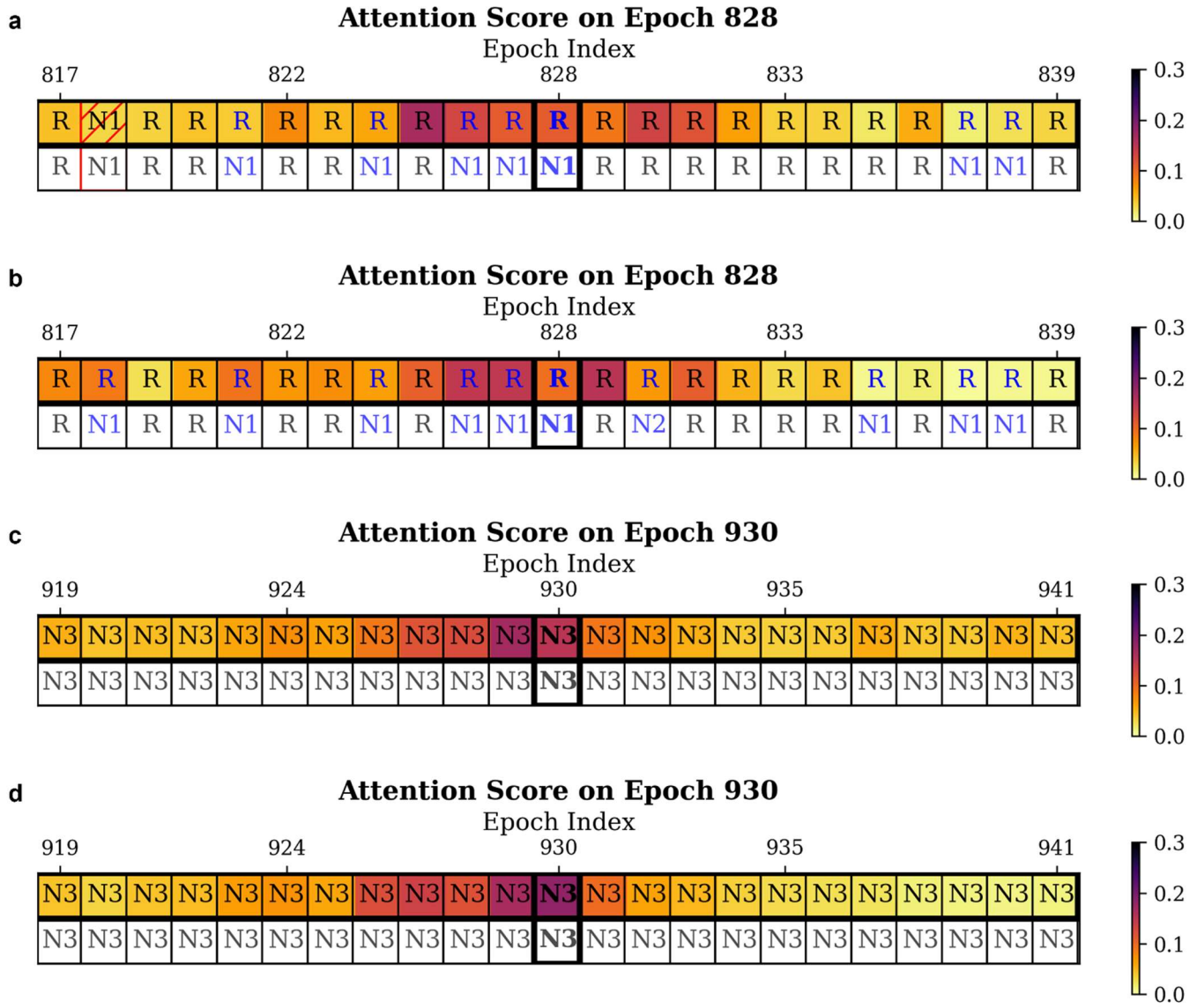

Figure 7. Inter-epoch interpretability of *DistillSleep*-S. Aggregated attention scores from *DistillSleep*-S and Baseline-S are visualized. The same sample epochs are analyzed as in the teacher model (see Figure 4). The hashed block in red indicates the incorrectly predicted epoch. Blue text indicates epochs where the stage 2 model corrected the stage 1 model's prediction. **a.** *DistillSleep*-S on REM epoch. **b.** Baseline-S on REM epoch. Comparing attention scores on REM epoch, *DistillSleep*-S demonstrates stronger attention scores on epoch 825 and 836 than Baseline-S, which resembles the teacher model. **c.** *DistillSleep*-S on N3 epoch. **d.** Baseline-S on N3 epoch. For N3 sequences, both *DistillSleep*-S and Baseline-S exhibit a relatively even distribution of attention scores, similar to the teacher model.

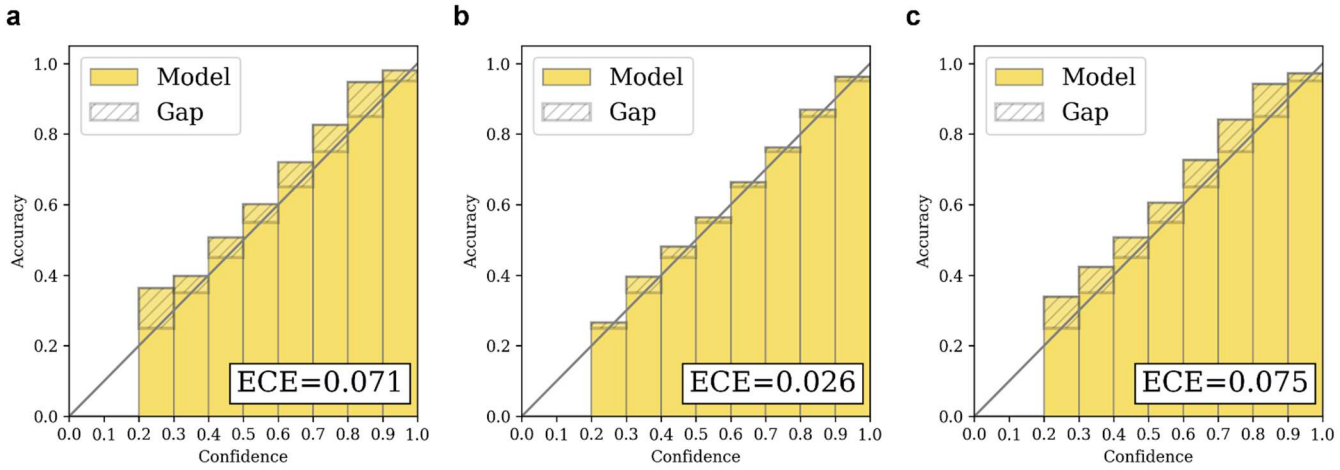

Figure 8. Model calibration results. The reliability diagrams of *DistillSleep*-T/S and Baseline-S are provided. ECE is provided at the right bottom of each subfigure. While *DistillSleep*-T and Baseline-S shows under-confidence, *DistillSleep*-S shows enhanced model calibration thanks to knowledge distillation. The calibration result is based on `PHY` dataset. **a.** Calibration results of *DistillSleep*-T. **b.** Calibration results of *DistillSleep*-S. **c.** Calibration results of Baseline-S.

Tables

Table 1. Predictive performance of the teacher model (*DistillSleep*-T). The best approach in each dataset per metric is marked as **bold**, and the second best approach is marked as underline. (*: U-Sleep [20] reports only performance on EEG-EOG input and is not directly comparable to *DistillSleep*-T.)

Dataset	Method	EEG				
		MF1	Acc.	Kappa	Sens.	Spec.
PHY	***DistillSleep*-T**	**78.9**	80.4	0.733	**79.3**	**94.7**
	***DistillSleep*-T (stage 1)**	75.6	77.6	0.696	76.4	94
	SleePyCo [24]	**78.9**	**80.9**	**0.737**	-	-
	XSleepNet [22]	78.6	80.3	0.732	78.7	94.6
	U-Time [21]	77.0	-	-	-	-
	U-Sleep* [20]	79.0	-	-	-	-
SHHS1	***DistillSleep*-T**	**81.1**	86.8	0.814	**82.1**	96.3
	***DistillSleep*-T (stage 1)**	74.0	82.7	0.754	73.9	95.1
	SleePyCo [24]	80.7	**87.9**	**0.830**	-	-
	SleepTransformer [23]	80.1	87.7	0.828	78.7	**96.5**
	XSleepNet [22]	80.7	87.6	0.826	79.7	**96.5**
	U-Sleep* [20]	80.0	-	-	-	-
	IITNet [19]	79.8	86.7	0.81	-	-
DCSM	***DistillSleep*-T**	**81.2**	**91.4**	**0.853**	**81.4**	**97.6**
	***DistillSleep*-T (stage 1)**	77.6	89.6	0.821	77.8	97
	U-Time [21]	79.0	-	-	-	-
	U-Sleep* [20]	81.0	-	-	-	-
KISS	***DistillSleep*-T**	**80.0**	**80.3**	**0.745**	**80.6**	**94.9**
	***DistillSleep*-T (stage 1)**	75.6	76.7	0.698	75.9	94.0
	SleepTransformer [23]	77.2	77.8	0.711	76.8	94.1

Table 2. On-device feasibility test. On edge devices, the combination of knowledge distillation and quantization reduces the e2e latency by up to 162x (Raspberry pi 4B), while preserving the Macro-F1 score at competitive level.

Device	Processors	Model	Framework	Precision	End-to-end Latency (ms)				Memory footprint (MB)
					Preprocessing	Stage 1	Stage 2	**Total**	
Workstation	CPU + GPU	*DistillSleep*-T	Tensorflow	FLOAT32	0.2	46.8	163.5	**210.5**	1109
	CPU	*DistillSleep*-S	tflite	INT8	0.1	2.6	0.6	**3.3**	73
Jetson orin nano	CPU + mGPU	*DistillSleep*-T	Tensorflow	FLOAT32	0.4	96.3	473.1	**569.8**	1191
	CPU	*DistillSleep*-S	tflite	INT8	0.4	3.7	1.4	**5.5**	58
Raspberry pi 4B	CPU	*DistillSleep*-T	Tensorflow	FLOAT32	0.9	342.5	1130.5	**1473.9**	491
	CPU	*DistillSleep*-S	tflite	INT8	0.9	5.7	2.5	**9.1**	52
Coral dev board	CPU + mNPU	*DistillSleep*-S	tflite	INT8	1.3	4.1	3.2	**8.6**	59

* FLOAT32: 32-bit floating point precision, INT8: 8-bit integer precision, mGPU: mobile Graphical Processing Unit, mNPU: mobile Neural Processing Unit.

Table 3. Improvement of *DistillSleep*-S's predictive power from knowledge distillation. The predictive performance for internal teacher and external teacher are presented separately. Both the Macro-F1 score and class-wise F1 scores are reported, with the increase in Macro-F1 score indicated in blue parentheses. Confusion matrices are provided in Supplementary Materials.

Teacher Type	Dataset	Method	EEG					
			MF1	W	N1	N2	N3	REM
Internal Teacher	PHY	*DistillSleep*-T	78.9	83.6	60.5	84.6	80.4	85.5
		***DistillSleep*-S**	**77.9 (+2.1)**	**82.6 (+1.1)**	**59.1 (+3.4)**	**84.1 (+1.0)**	**80.2 (+2.4)**	**83.5 (+2.5)**
		Baseline-S	75.8	81.5	55.7	83.1	77.8	81.0
	SHHS1	*DistillSleep*-T	81.1	91.0	53.7	88.4	83.5	88.8
		***DistillSleep*-S**	**79.7 (+2.3)**	**89.9 (+1.2)**	**51.4 (+2.6)**	**87.6 (+1.5)**	**82.4 (+4.3)**	**87.0 (+1.7)**
		Baseline-S	77.4	88.7	48.8	86.1	78.1	85.3
	DCSM	*DistillSleep*-T	81.2	97.5	50.3	84.7	84.9	88.9
		***DistillSleep*-S**	**78.8 (+3.4)**	**96.9 (+0.5)**	**46.1 (+9.0)**	**82.8 (+3.3)**	**84.1 (+1.5)**	**84.2 (+2.8)**
		Baseline-S	75.4	96.4	37.1	79.5	82.6	81.4
	KISS	*DistillSleep*-T	80.0	86.6	63.5	80.1	81.1	88.8
		***DistillSleep*-S**	**77.8 (+1.9)**	**84.6 (+1.8)**	**59.9 (+3.1)**	**79.0 (+1.0)**	**80.0 (+0.9)**	**85.4 (+2.7)**
		Baseline-S	75.9	82.8	56.8	78.0	79.1	82.7
External Teacher	SleepEDF-78	***DistillSleep*-S**	**76.3 (+5.0)**	**91.8 (+0.9)**	**52.2 (+9.7)**	**84.1 (+2.3)**	**70.8 (+0.7)**	**82.4 (+11.1)**
		Baseline-S	71.3	90.9	42.5	81.8	70.1	71.3
	ISRUC-SG1	***DistillSleep*-S**	**74.7 (+4.6)**	**84.6 (+3.7)**	**48.6 (+9.1)**	**76.0 (+2.5)**	**86.7 (+2.3)**	**77.3 (+4.9)**
		Baseline-S	70.1	80.9	39.5	73.5	84.4	72.4
	ISRUC-SG2	***DistillSleep*-S**	**71.1 (+6.8)**	**80.1 (+10.5)**	**47.5 (+8.6)**	**71.4 (+5.3)**	**85.5 (+2.2)**	**70.8 (+7.2)**
		Baseline-S	64.3	69.6	38.9	66.1	83.3	63.6
	ISRUC-SG3	***DistillSleep*-S**	**77.1 (+2.1)**	**89.0 (+0.9)**	**52.6 (+7.3)**	**78.9 (-0.1)**	**88.1 (+0.5)**	**76.6 (+1.8)**
		Baseline-S	75.0	88.1	45.3	79	87.6	74.8

Table 4. Effect of quantization. Predictive performance and model size comparison of *DistillSleep*-S by different frameworks and quantization approaches are presented. Evaluated on `PHY`**.**

Framework	Quantization Approach	Precision	Size (MB)	Macro-F1 (%)	Diff.
TensorFlow	-	FLOAT32	3.019	77.9	-
TensorFlow Lite	-	FLOAT32	0.441	77.9	-
TensorFlow Lite	QAT with KD	INT8	0.191	77.7	0.2
	QAT			76.6	1.3
	PTQ			46.7	31.2

* FLOAT32: 32-bit floating point precision, INT8: 8-bit integer precision, KD: Knowledge Distillation, QAT: Quantization Aware Training, PTQ: Post Training Quantization.

Table 5. Generalizability test. External teacher trained on four large datasets is evaluated on the unseen SleepEDF-78 and ISRUC datasets. Performance is shown in zero-shot setting and after applying two test-time adaptation methods to improve generalizability. Results from a fully finetuned model are included as an upper-bound reference. All values are Macro-F1 scores.

Dataset	External validation	Test time BN adaptation [59]	TENT [60]	Full finetuning (Upper bound)
SleepEDF-78	56.5	66.2	66.9	75.9
ISRUC-SG1	65.3	74.8	75.7	78.6
ISRUC-SG2	55.9	70.1	71.3	73
ISRUC-SG3	67.8	75.8	76.3	81.1

Table 6. Robustness to EEG lead choice. This table presents a robustness test comparing the performance of *DistillSleep*-T/S and Baseline-S on two additional EEG leads (F3-M2 and O1-M2) against the C4-M1 lead used in the main study.

Teacher/Student	EEG lead	MF1	Acc.	Kappa	Sens.	Spec.
DistillSleep-T	C4-M1	78.9	80.4	0.733	79.3	94.7
	F3-M2	79.5	81	0.742	79.9	94.8
	O1-M2	77	78.8	0.711	77.3	94.2
DistillSleep-S	C4-M1	77.9	79.5	0.721	78.3	94.4
	F3-M2	79	80.7	0.735	79	94.7
	O1-M2	76.1	77.9	0.697	75.9	93.9
Baseline-S	C4-M1	75.8	78	0.699	76	93.9
	F3-M2	77.4	79.4	0.718	77.8	94.4
	O1-M2	73.9	76	0.673	74.4	93.4