

<https://doi.org/10.1038/s41746-024-01378-0>

# Explainable vision transformer for automatic visual sleep staging on multimodal PSG signals



Hyojin Lee<sup>1,8</sup>, You Rim Choi<sup>1,8</sup>, Hyun Kyung Lee<sup>2,7,8</sup>, Jaemin Jeong<sup>3</sup>, Joopyo Hong<sup>1</sup>,  
Hyun-Woo Shin<sup>2,4,5,6,7</sup> ✉ & Hyung-Sin Kim<sup>1</sup> ✉

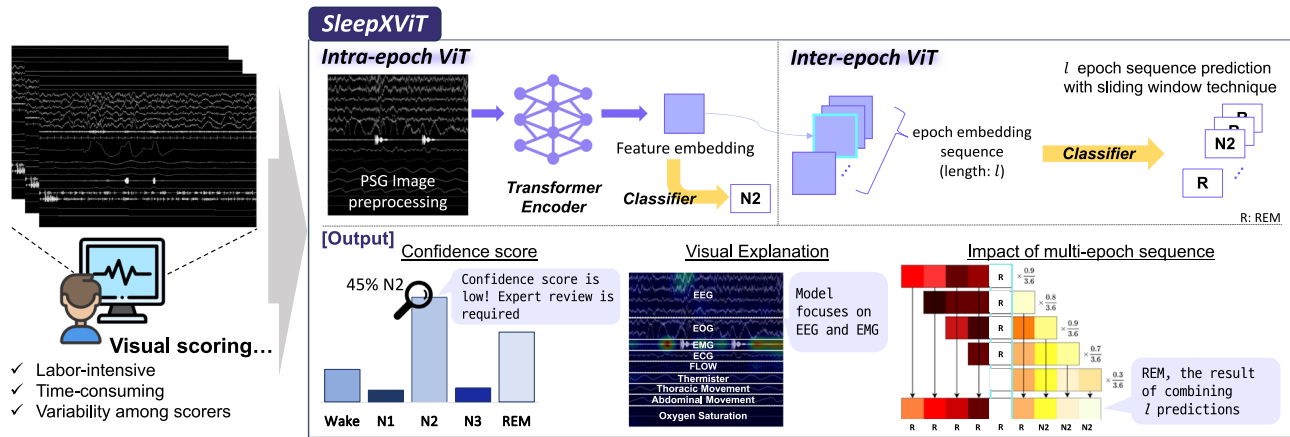
Polysomnography (PSG) is crucial for diagnosing sleep disorders, but manual scoring of PSG is time-consuming and subjective, leading to high variability. While machine-learning models have improved PSG scoring, their clinical use is hindered by the ‘black-box’ nature. In this study, we present *SleepXViT*, an automatic sleep staging system using Vision Transformer (ViT) that provides intuitive, consistent explanations by mimicking human ‘visual scoring’. Tested on KISS—a PSG image dataset from 7745 patients across four hospitals—*SleepXViT* achieved a Macro F1 score of 81.94%, outperforming baseline models and showing robust performances on public datasets SHHS1 and SHHS2. Furthermore, *SleepXViT* offers well-calibrated confidence scores, enabling expert review for low-confidence predictions, alongside high-resolution heatmaps highlighting essential features and relevance scores for adjacent epochs’ influence on sleep stage predictions. Together, these explanations reinforce the scoring consistency of *SleepXViT*, making it both reliable and interpretable, thereby facilitating the synergy between the AI model and human scorers in clinical settings.

Sleep is a vital biological process crucial for maintaining overall health, influencing energy restoration<sup>1</sup>, immune function<sup>2</sup>, and cognitive performance<sup>3,4</sup>. It is also linked to various health conditions, including cardiovascular<sup>5</sup> and metabolic diseases<sup>6,7</sup>, and depression<sup>6</sup>. Thus, monitoring sleep and analyzing sleep patterns have been longstanding interests in the medical field<sup>8,9</sup>. The gold standard for clinical sleep monitoring is polysomnography (PSG)<sup>10</sup>, which comprehensively measures various physiological parameters, including electroencephalography (EEG), electrocardiography (ECG), electrooculography (EOG), electromyography (EMG), and respiratory activity. Following PSG, skilled technicians perform sleep scoring that involves classifying each 30-second sleep period (1 epoch) into one of five stages according to the American Academy of Sleep Medicine (AASM) scoring rules: wake (W), rapid eye movement (REM) sleep, and three non-REM (NREM) sleep stages—N1, N2, and N3<sup>11,12</sup>. This classification is essential for evaluating sleep quality and diagnosing related disorders<sup>13</sup>. However, the PSG scoring process is labor-intensive, time-consuming, and prone to variability among scorers<sup>14</sup>, which introduces significant limitations.

To address these limitations, significant strides have been made toward automating sleep scoring<sup>15–17</sup>. Recent advancements in deep learning (DL) have led to the development of DL-based methods that achieve clinically acceptable accuracy. A predominant approach leverages single-channel EEG data, focusing on the frequency domain to capture distinct characteristics associated with each sleep stage. To account for the continuity of sleep stages, models<sup>18–21</sup> have been developed to process EEG across a multi-epoch sequences as input. However, information from a single channel remains limited, prompting some studies<sup>22,23</sup> to integrate additional channels, such as EOG or EMG, to improve performance. In clinical practice, however, sleep experts classify sleep stages by visually assessing all multimodal PSG data displayed on a monitor, including not only EEG, EOG, and EMG signals but also ECG and respiratory signals. Consequently, a gap still exists between the information traditionally used by human scorers and the data leveraged by these models for sleep stage classification.

Another major limitation of existing methods is the ‘black box’ nature of DL algorithms, which hinders clinical adoption due to their lack of interpretability and transparency<sup>24–26</sup>. Sleep stage scoring is complex and

<sup>1</sup>Graduate School of Data Science, Seoul National University, Seoul, Republic of Korea. <sup>2</sup>Obstructive Upper Airway Research (OUAR) Laboratory, Department of Pharmacology, Seoul National University College of Medicine, Seoul, Republic of Korea. <sup>3</sup>Department of Computer Engineering, School of Software, Hallym University, Chuncheon, Republic of Korea. <sup>4</sup>OUAR LaB, Inc, Seoul, Republic of Korea. <sup>5</sup>Department of Otorhinolaryngology-Head and Neck Surgery, Seoul National University Hospital, Seoul, Republic of Korea. <sup>6</sup>Sensory Organ Research Institute, Seoul National University Medical Research Center, Seoul, Republic of Korea. <sup>7</sup>Present address: Department of Biomedical Sciences, Seoul National University Graduate School, Seoul, Republic of Korea. <sup>8</sup>These authors contributed equally: Hyojin Lee, You Rim Choi, Hyun Kyung Lee. ✉e-mail: [charlie@snu.ac.kr](mailto:charlie@snu.ac.kr); [hyungkim@snu.ac.kr](mailto:hyungkim@snu.ac.kr)



**Fig. 1 | SleepXViT Overview.** Raw PSG signal values are converted into standardized PSG images. The *Intra-epoch ViT* learns features from these images and includes a sleep stage classifier. The *Inter-epoch ViT* processes sequences of feature embeddings from the *Intra-epoch ViT* for  $l$  consecutive epochs, analyzing the relationships between epochs within the sequence to predict sleep stages for

each epoch. *SleepXViT* not only predicts sleep stages but also provides explanations for them: 1) a confidence score indicating the certainty of the prediction, 2) a visual explanation by heatmap that highlights crucial parts of the PSG image for decision-making, and 3) the impact of multi-epoch sequence on the prediction.

often results in variability among scorers, with inter-rater reliability around 82%<sup>27–29</sup>. Consequently, scorers must verify their decisions against each other to ensure consistency in criteria. This requirement also applies to DL methods, which must provide evidence or clues to support their decisions. Studies such as SleepTransformer<sup>18</sup> have introduced DL models with inference mechanisms that can identify and highlight parts of the data as scoring clues<sup>18,30,31</sup>. However, these models utilize only specific raw signal values, such as EEG or EOG, analyzing them within the time-frequency domain. Therefore, it becomes challenging for humans scorers, who perform ‘comprehensive visual scoring’ across multiple signal types, to intuitively understand and accept the results presented by these models, which focus on limited signal sources.

In light of these challenges, the introduction of the Korea Image-based Sleep Study (KISS) dataset into sleep research<sup>32</sup> offers promising solutions to overcome these limitations. This dataset facilitates the standardization and preprocessing of PSG raw data, enabling a visual approach to sleep scoring that reflects human operations (Supplementary Note 1 and Supplementary Fig. 1). By utilizing the standardized pipeline used to generate this dataset, PSG data collected from various systems can be consistently transformed into a uniform image dataset composed of raw signals. Jeong et al. developed a sleep stage classification model based on Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM), which utilizes the Eigen-CAM method<sup>33</sup> to visually explain decisions by highlighting important parts in images through heatmaps. These heatmaps are generated from the feature map of the final convolutional layer, often representing broader, more generalized regions and resulting in coarse-grained depictions that do not comprehensively cover multiple areas. Additionally, while CNNs excel at capturing local features, their limited ability to grasp global context restricts the full utilization of all signal information within the PSG images. Consequently, these limitations still result in insufficient interpretability and fail to provide clear insights into the decision-making process.

To this end, we propose *SleepXViT*, a novel automatic sleep staging system designed to synergize with clinicians by ensuring it is reliable and understandable within a clinical setting. It mimics human ‘visual scoring’ by using a PSG image that includes all multimodal waveforms collected during the PSG as inputs, enabling intuitive understanding for humans. *SleepXViT* utilizes a Vision Transformer (ViT)<sup>34</sup>, as detailed in the overview shown in Fig. 1. The *Intra-epoch ViT* extracts features from a single epoch image and classifies it into specific sleep stages. Subsequently, the *Inter-epoch ViT* takes sequences of epochs processed by the *Intra-epoch ViT*, generates sleep stage sequences, and aggregates multiple sequence predictions to accurately

determine the sleep stage for each epoch. *SleepXViT* not only achieves state-of-the-art performance with well-calibrated confidence scores but also ensures intuitive and consistent explainability across different sleep stages.

Our contributions can be summarized as follows:

- *SleepXViT* achieves superior classification performance and provides reliable confidence scores for predictions, enabling experts to thoroughly review results predicted with low confidence. This significantly increases both reliability and usability in clinical settings.
- *SleepXViT* provides visual explanations through high-resolution, epoch-by-epoch heatmaps that highlight the parts of images the model focuses on for its classification predictions. It enables human experts to verify whether the model’s staging rationale can be explained using the AASM rules they employ.
- Additionally, it analyzes the impact of multi-epoch sequences, mirroring the way human experts consider the temporal context in their staging decisions, thus providing a more comprehensive evaluation.
- We assess the consistency with which the model applies its staging criteria, ensuring that the algorithm’s internal workings are clearly explainable.

## Results

### Baselines and Datasets

For our performance comparison, we used two explainable sleep staging models as baselines: the model developed by Jeong et al.<sup>32</sup> and SleepTransformer<sup>18</sup>. Like *SleepXViT*, Jeong et al. utilized standardized PSG images with multimodal biosignals as inputs, employing a CNN with a bidirectional LSTM (Bi-LSTM) and highlighting regions of focus using the Eigen-CAM<sup>33</sup> method. SleepTransformer, on the other hand, used single-channel EEG data, converting 30-second EEG epochs into time-frequency images as inputs. It employed a Transformer<sup>35</sup> to predict sleep stages from multi-epoch sequences, providing both attention heatmaps and prediction confidence.

We utilized two primary datasets, with a significant focus on the KISS dataset<sup>32</sup>, a standardized image-based PSG dataset sourced from AI Hub<sup>36</sup>. This dataset comprises over 7.62 million images from retrospectively collected 10,253 PSG records, representing a wide demographic range and including various physiological signals such as EEG, EOG, and ECG. The images, capturing 30-second intervals of PSG results from 2013 to 2020, have been collected across four different hospitals using two distinct PSG devices, highlighting the dataset’s extensive coverage and standardization. The KISS dataset is particularly notable for its broad signal range and the demographic diversity of its subjects, predominantly male, mirroring the typical patient population for sleep disorders<sup>37</sup>, including conditions like

obstructive sleep apnea (OSA). To ensure consistency across different systems, Jeong et al. established a *standardized* preprocessing pipeline that converts heterogeneous biosignals into a single image format, maintaining applicability across varying software, signal orders, or screen dimensions. For our study, we selected 7,745 PSG records from this dataset, dividing them into training, validation, and testing datasets with patient-based split ratios of 80/10/10, as described by Jeong et al. In applying SleepTransformer to this dataset, raw EEG signal values were used, instead of images.

The second dataset employed is the Sleep Heart Health Study (SHHS), which examines the link between sleep disorders and cardiovascular health<sup>38,39</sup> (Supplementary Note 2). The dataset is divided into visit 1 (SHHS1) and visit 2 (SHHS2) based on the exam cycle. We used 5,452 PSG records from the SHHS1 dataset and 2,651 PSG records from the SHHS2 dataset, converting them to 5.87 million and 3.19 million standardized images, respectively, following the standardized preprocessing pipeline<sup>32</sup>. For SHHS (Visit1), to ensure comparability, we aligned our data split with that used by SleepTransformer, assigning 70% for training and 30% for testing, while holding out 100 subjects from the training set for validation. For SHHS (Visit2), we set aside 10% for testing, dividing the remaining data into 80% for training and 20% for validation, mirroring the experimental setup by Jeong et al. to facilitate direct comparison. Given the fewer available signal channels compared to KISS, we adapted the images by employing a method from Jeong et al., which involves duplicating similar signals where channels are missing. We also used blank placeholders to maintain consistency in the dataset's structure<sup>32</sup>. This approach allows us to manage the disparity in channel availability effectively. Detailed information on the datasets and the preprocessing steps involved in converting the data into images can be found in the Methods section and Supplementary Fig. 2.

Classification Performance and Confidence

Table 1 presents a performance comparison of *SleepXViT* with other baseline models across various datasets, using Accuracy, Macro F1 Score (MF1), Cohen's Kappa ( $\kappa$ ), and Weighted F1 Score (WF1) as evaluation metrics. Given the class imbalance in sleep stages, MF1 serves as the primary metric to ensure balanced assessments across all classes, reflecting both precision and recall. To enhance clinical relevance and inter-rater reliability, we have also included  $\kappa$  in the evaluation, allowing for a clearer comparison with previous sleep staging studies. To complement the quantitative assessment presented in Table 1, Supplementary Figs. 3, 4, and 5 illustrate hypnograms corresponding to the highest, median, and lowest F1 scores observed in KISS, SHHS1, and SHH2, respectively. These visualizations provide a direct representation of how well *SleepXViT* capture the dynamics of sleep stages, highlighting the models' performance in practical, clinical contexts.

In tests on 775 subjects from the KISS dataset, *SleepXViT* achieves the highest overall MF1 at 81.94, demonstrating superior performance across nearly all class-wise MF1 metrics. In the SHHS1 dataset, *SleepXViT* demonstrated performance comparable to that of SleepTransformer. While it exhibited a marginally lower MF1 than SleepTransformer, it achieved higher values in other metrics such as  $\kappa$ , Accuracy, and WF1, indicating its overall robustness. Furthermore, on the SHHS2 dataset, *SleepXViT* significantly outperformed all other baseline models, achieving superior results across four key overall metrics by a considerable margin. Notably, *SleepXViT*'s  $\kappa$  is consistently higher than that of other interpretable sleep-staging models across all three datasets, underscoring its robust performance and reliability. Additionally, we report the performance of using only the *Intra-epoch ViT* component of *SleepXViT* on the KISS dataset, which slightly underperforms relative to Jeong et al. that considers epoch sequences. Applying the *Inter-epoch ViT*, which utilizes multi-epoch sequences and considers adjacent epochs, leads to a 1.84 percentage point increase in MF1. The impact of varying input epoch sequence lengths on performance is detailed in Table 2; an input length of 10 achieves the highest MF1, leading us to adopt this length for all experiments.

*SleepXViT* demonstrates strong performance on the test dataset but is not perfect. Therefore, for practical use in clinical settings, it is crucial to report the confidence score, which indicates the likelihood of the model's predictions being accurate. This enables human experts to assess whether they can rely on the model's predictions. Previous studies have explored confidence or uncertainty estimates in sleep staging models to enhance reliability and interpretability. For instance, Fiorillo et al.<sup>40</sup> utilized Monte Carlo dropout to estimate uncertainty through sub-network predictions, while Van Gorp et al.<sup>41</sup> generated multiple hypnograms to model inter-rater disagreement with a conditional generative model. However, both approaches require multiple model runs, leading to increased evaluation time. Similarly, Heremans et al.<sup>42</sup> proposed an uncertainty-guided pipeline but relied on a single-scoring labeled dataset, potentially overlooking inter-rater variability. In contrast, we opted to use softmax values as a confidence metric, allowing for a straightforward and practical assessment of the model's predictive accuracy. The softmax function, typically employed in the output layer of deep learning models, converts the model's outputs into probabilities that represent the likelihood of each class. By utilizing softmax values, we can efficiently evaluate the results of *SleepXViT* and enable human experts to make informed decisions. When the model predicts with low confidence, manual verification and adjustments become necessary.

Table 1 | The performance comparison of *SleepXViT* with other baselines across various datasets

Database	System	Input Channels	Overall Metrics (%)				Class-wise MF1 (%)				
			Acc.	$\kappa$	MF1	WF1	W	N1	N2	N3	REM
KISS	<i>SleepXViT</i>	Image	<b>82.49</b>	<b>77.27</b>	<b>81.94</b>	<b>82.52</b>	88.91	64.36	82.44	82.13	90.20
	<i>SleepXViT</i> (Intra-Epoch ViT)	Image	80.84	75.10	80.10	80.91	88.29	63.40	80.70	81.05	87.07
	Jeong et al. <sup>32</sup>	Image	<b>82.10</b>	<b>76.61</b>	<b>81.17</b>	<b>82.02</b>	88.95	64.30	81.99	81.29	89.30
	Jeong et al. <sup>32</sup>	Signal	80.91	75.22	80.36	81.09	88.20	64.41	80.48	79.91	88.81
	SleepTransformer <sup>18</sup>	Signal	77.84	71.05	77.17	77.82	83.94	55.89	78.53	80.57	86.93
SHHS1	<i>SleepXViT</i>	Image	<b>88.07</b>	<b>83.20</b>	<b>80.18</b>	<b>87.90</b>	93.29	45.34	88.57	83.50	90.18
	Jeong et al.	Image	87.02	81.65	78.53	<b>86.76</b>	92.63	42.24	87.68	80.93	89.16
	SleepTransformer	Signal	<b>87.76</b>	<b>82.89</b>	<b>80.47</b>	83.39	92.24	47.68	88.39	85.19	88.83
SHHS2	<i>SleepXViT</i>	Image	<b>90.51</b>	<b>86.43</b>	<b>81.55</b>	<b>90.31</b>	91.61	45.44	89.87	84.83	91.99
	Jeong et al.	Image	83.05	75.71	70.94	82.63	91.44	24.86	81.52	79.22	77.69
	SleepTransformer	Signal	<b>88.89</b>	<b>84.17</b>	<b>79.58</b>	<b>88.54</b>	94.00	41.81	89.02	84.91	88.16

Accuracy, Cohen's Kappa ( $\kappa$ ), Macro F1 Score (MF1), and Weighted F1 Score (WF1) are used as evaluation metrics. Considering the class imbalance in sleep stages, MF1 is employed as the primary metric to ensure balanced assessments across all classes and to reflect both precision and recall. The best result is marked in bold underline, and the second-best result is marked in bold.

**Table 2 | Performance variation of the *Inter-epoch ViT* with changes in the length of the multi-epoch sequence (*l*)**

Database	Input Length ( <i>l</i> )	Overall metrics			Class-wise MF1				
		Acc.	MF1	WF1	W	N1	N2	N3	REM
KISS	5	82.51	81.61	82.40	88.91	64.36	82.44	82.13	90.20
	10	82.49	<b>81.94</b>	82.52	88.76	65.95	81.94	82.70	90.35
	15	82.47	81.75	82.43	88.79	65.05	82.16	82.52	90.26
	20	82.40	81.83	82.49	88.85	65.69	82.11	82.25	90.27
	25	82.44	81.67	82.40	88.85	64.58	82.27	82.36	90.32

A sequence length of 10 shows the highest overall MF1, and is therefore used for all experiments.

With this understanding, ‘can we then trust the confidence scores provided by the model?’ Evaluating their reliability is essential and is achieved through model calibration<sup>43</sup>. This process assesses how well the model’s predicted confidence scores align with actual outcomes. A model is considered perfectly calibrated when its confidence reflects true accuracy. For example, if the model predicts that a sample belongs to a specific class with 70% confidence, ideally, there should be a 70% probability that the sample indeed belongs to that class. Avoiding over-confidence, where the model exhibits higher confidence than its actual accuracy, is crucial when utilizing the confidence scores in practical medical decision-making<sup>44</sup>. The reliability diagram visualizes model calibration by plotting confidence intervals (bins) on the x-axis against the prediction accuracy for those intervals on the y-axis. A perfectly calibrated model would have bars in the diagram aligning with the  $y = x$  line. The reliability diagram for *SleepXViT*, as shown in Fig. 2a, demonstrates that the bars are close to the  $y = x$  line, indicating that *SleepXViT* is well-calibrated. In the reliability diagram, the ‘Gap’ represents the difference between the average confidence and accuracy within specific confidence intervals. The *Intra-epoch ViT* tends to show either nearly equal or slightly lower average confidence compared to its accuracy, while the *Inter-epoch ViT* exhibits a slight tendency towards overconfidence. Figure 2b displays the distribution of confidence across the entire test dataset, showing a higher confidence distribution when using the *Inter-epoch ViT*.

The distribution of class-wise confidence scores, as illustrated in Fig. 2c, underscores the importance of monitoring these scores and demonstrates their practical utility. When analyzed alongside the normalized confusion matrix in Fig. 2d, there is a pronounced trend where classes with higher average confidence tend to have higher accuracy. Notably, for the N1 class, which exhibits lower accuracy, confidence scores are relatively low and evenly spread between 0.4 and 0.9, with few exceeding 0.9. This pattern suggests that *SleepXViT* is capable of identifying epochs likely to be misclassified, based on their confidence scores. Figure 2e shows the distribution of confidence for cases where the results predicted by the *Intra-epoch ViT* are modified by the *Inter-epoch ViT*. Interestingly, in these samples, the *Intra-epoch ViT* predicts with an average confidence of 0.54, which is considerably overconfident given that the actual accuracy is only 0.35. When the *Inter-epoch ViT* is applied, the accuracy increases by 0.2, and the average confidence increases by 0.065, reducing the gap between these metrics.

### Visual explanations

Visual explanations transparently illustrate how models make decisions by highlighting the input features with the most significant impact. This clarity enhances user trust and supports necessary modifications or enhancements to the model’s decisions<sup>45</sup>. In medical imaging analysis, heatmaps are commonly used to display local relevance and emphasize the key regions of an input image that greatly affect the outcomes. Human experts can review the heatmap provided by the model to assess whether the model’s reasoning aligns with their decision-making criteria or is convincingly accurate.

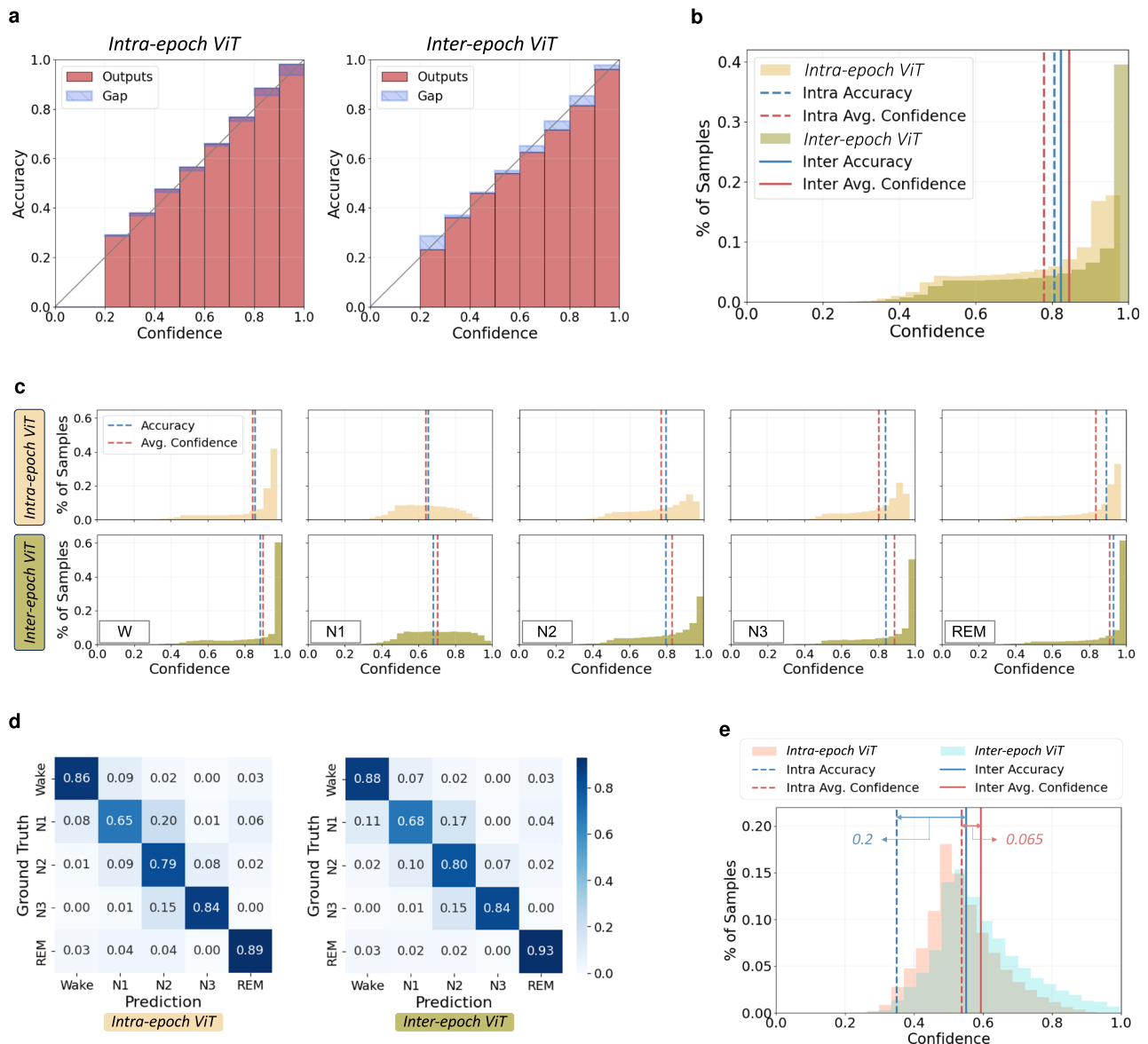
Various methods have been developed to generate these heatmaps specifically for CNNs<sup>33,45–47</sup>. Grad-CAM<sup>47</sup>, a widely utilized technique, uses the feature map activations from convolutional layers weighted by the gradients for a particular output class to visualize importance on a class-specific basis. However, as Grad-CAM relies solely on gradients from the

deepest layers, the resultant heatmaps, upscaled from low-resolution maps, are often coarse. Similarly, Jeong et al. employed Eigen-CAM<sup>33</sup>, which also generates heatmaps from the deepest feature maps and encounters the same resolution issues. Another approach, Layer-wise Relevance Propagation (LRP)<sup>48</sup>, backpropagates signals from the output to the input to analyze each pixel’s contribution to the final decision in detail. LRP provides high-resolution heatmaps that precisely detail the relevance of each component of the input to the predictions, offering a more granular view than methods relying solely on deep-layer gradients.

We adopted a novel method<sup>49</sup> designed for Transformer networks, based on LRP, as the visual explanation method for *SleepXViT*. This method determines local relevance using the Deep Taylor Decomposition principle and subsequently propagates these relevancy scores through the layers. Transformer networks, which include attention layers and skip connections, present challenges for existing methods in this propagation. However, this approach effectively maintains total relevancy across the layers. We successfully applied this method to the *Intra-epoch ViT*, which extracts features from input PSG images. The heatmaps (relevance maps) created by this method visualize how each pixel in the input image contributes to the final decision, represented as a relevance score. This enables us to identify the key regions that play a critical role in the model’s decision-making process. Figure 3a shows the average heatmap generated by overlaying 10,000 maps for each class, revealing the tendencies of which signal areas influence the decisions for each class. Figure 3b presents representative samples that vividly emphasize the distinctive features of each class, clearly indicating which signal areas and characteristics are being focused on. For additional samples from other patients, see Supplementary Figs. 6, 7, and 8.

In the KISS dataset, EEG is recorded from four locations: two from central electrodes displayed at the top (C3A2 and C4A1) and two from occipital electrodes displayed at the bottom (O2A1 and O1A2). In the Wake stage, Fig. 3a shows that, on average, the EEG regions from these four channels are accentuated throughout the entire epoch. The EOG, although subtle, is also more pronounced compared to other regions, indicating eye movements as seen in Fig. 3b. For N1, the EEG is broadly observed, with a higher emphasis on measurements from the central region. Additionally, there is attention on EOG and respiratory-related Flow signals. In N2, only specific parts of the EEG are slightly underscored on average. Figure 3b illustrates that in N2, the model concentrates on much more localized EEG regions compared to other sleep stages, resulting in a lower average relevance score when these heatmaps are averaged. A detailed examination of these localized regions in N2 reveals *sleep spindles* and *K-complexes*, which are characteristic features of this stage. For N3, the heatmaps show that all EEG channels and the EOG are prominently marked across the entire epoch. In the REM stage, although EOG and EMG signals are highlighted, there is a predominant focus on the Flow signal, with other respiratory-related movements also being significantly noted. From the analysis of the heatmaps in Fig. 3, it can be interpreted that the Wake, N1, and N2 stages are consistent with the AASM rules used by human experts for sleep staging. However, N3 and REM stages require further discussion. The relationship between the heatmaps for each sleep stage and the AASM guidelines, as well as a detailed discussion on N3 and REM, will be elaborated in the Discussion section.





**Fig. 2 | Analysis of confidence scores in SleepXViT.** This analysis focuses on evaluating the confidence scores generated by the *Intra-epoch ViT* and *Inter-epoch ViT* on the KISS test set. **a** Reliability diagram illustrating the gap, which represents the difference between average confidence and accuracy within specific confidence intervals. **b** Distribution of confidence scores for the predictions. **c** Distribution of

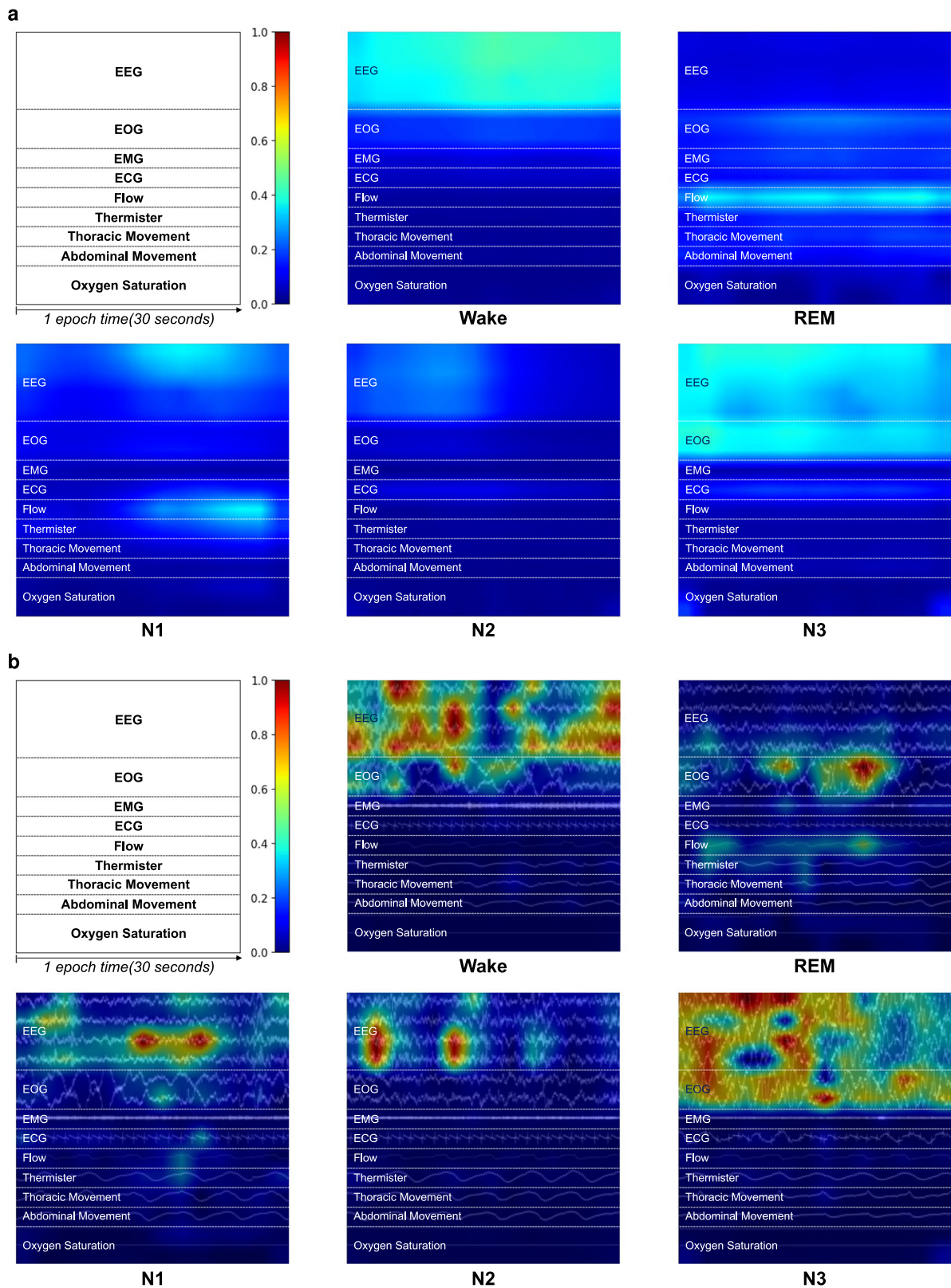
confidence scores from *Intra-epoch ViT* and *Inter-epoch ViT*, categorized by each sleep stage. **d** Class-wise confusion matrix with normalization. **e** Distribution of confidence scores for cases where the results predicted by the *Intra-epoch ViT* are modified by the *Inter-epoch ViT*.

### Impact of Multi-epoch Sequences

We have observed that the *Inter-epoch ViT* significantly enhances classification performance within SleepXViT. ‘How does the *Inter-epoch ViT* manage to bring about such performance improvements?’ Fig. 4a details the inference process and explainability analysis for the *Inter-epoch ViT*. It processes a sequence of multiple epochs using a sliding-window approach with a 1-epoch interval. Within each window, attention mechanisms consider the epochs surrounding the target epoch, allowing for adjustments in its predictions. As the window slides, the epochs adjacent to the target epoch change, potentially altering its outcome. Even if the specific predictions remain the same, the probabilities associated with those predictions may vary. This process results in multiple predictions for a single target epoch, and we combined these predictions by summing their softmax values, selecting the class with the highest total as the final prediction. This ensemble effect of predictions across overlapping epoch sequences helps to enhance prediction accuracy. Relevance scores indicate how much other epochs in the input sequence are referenced when predicting the target

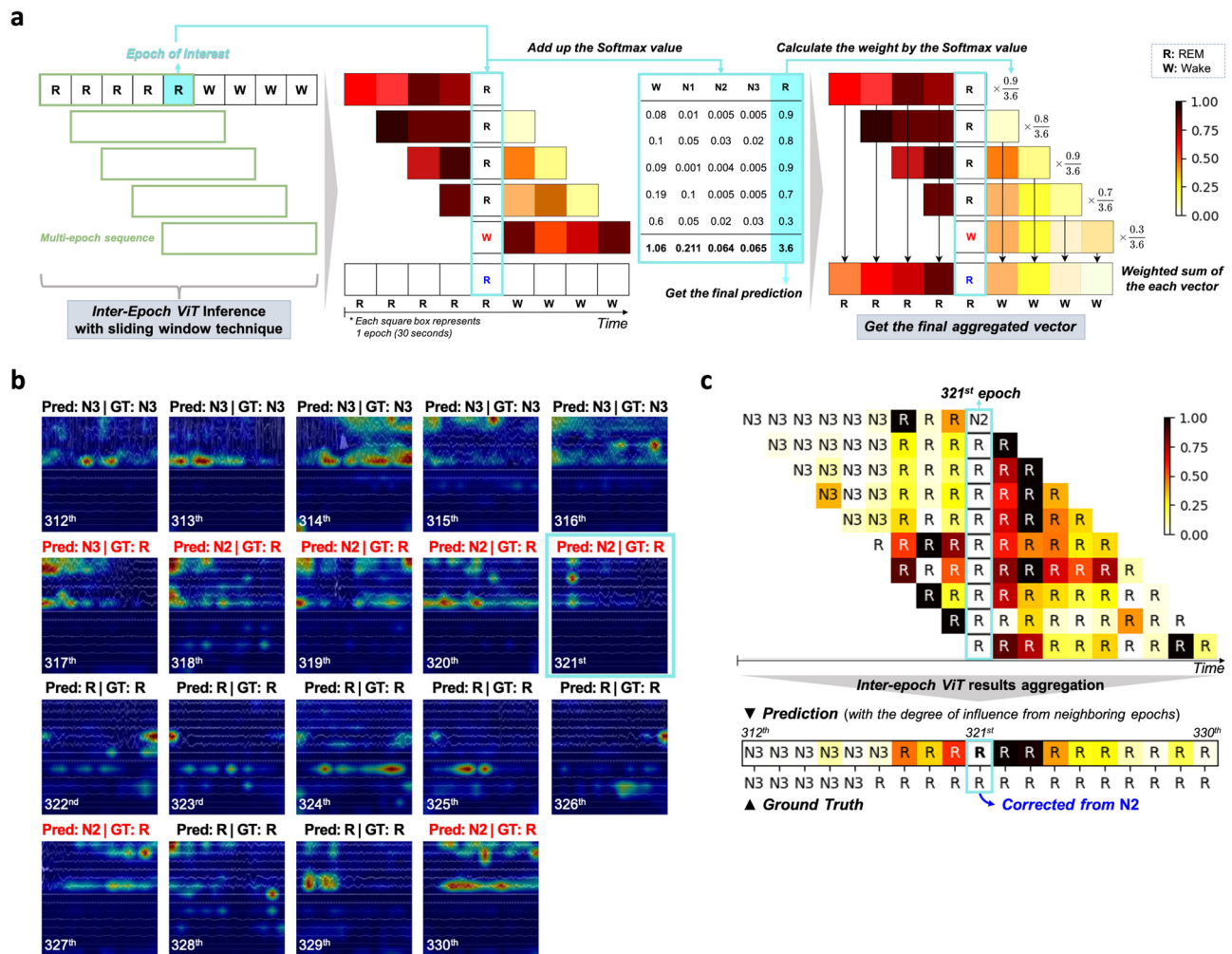
epoch. To compute the relevance scores between epochs, we utilized the methodology applied to the *Intra-epoch ViT* as described in the ‘Visual Explanations’ section. These scores are then aggregated using softmax values, which helps normalize and weigh the influence of each epoch, producing a final, weighted relevance score. The color scale in Fig. 4a indicates the level of relevance, with higher values representing greater relevance to the target epoch.

Figure 4b presents the ground truth, prediction results of the *Intra-epoch ViT*, and corresponding input images with overlaid heatmaps for a patient’s epochs from the 312<sup>th</sup> to the 330<sup>th</sup>. While the ground truth for epochs starting from the 317<sup>th</sup> is REM, seven out of fourteen epochs are incorrectly classified. The visual explanations for these misclassified epochs exhibit ambiguous features that could easily confuse even human scorers. It is important to note that REM scoring often involves considering the context of adjacent epochs, which may not exhibit distinct REM characteristics but are still classified as REM based on their surrounding epochs. This context-based approach is effectively mimicked by the *Inter-epoch ViT*,



**Fig. 3 | Visual explanations of SleepXViT on KISS dataset.** Heatmaps emphasize the key regions of the input image that significantly affect the outcomes. The colors represent the relevance of each pixel to the model's decision. **a** The average heatmap,

created by overlaying 10,000 maps for each class, demonstrates which signal parts influence the decisions for each class across numerous samples. **b** Representative samples that vividly highlight distinctive features characteristic of each class.



**Fig. 4 | Influence of the *Inter-epoch ViT*.** Visualizations demonstrate the influence and role of the *Inter-epoch ViT*. **a** Illustration of the inference process and explainability analysis for the *Inter-epoch ViT*. The color scale bar represents the relevance to the target epoch, with higher values indicating greater relevance. **b** For the target epoch (the 321<sup>st</sup>), this figure shows the ground truth and the prediction

results of the *Intra-epoch ViT*, alongside corresponding input images with overlaid heatmaps from epochs 312<sup>th</sup> to 330<sup>th</sup> of a patient. **c** Correction of predictions by the *Inter-epoch ViT*, which accurately classifies the 321<sup>st</sup> epoch as REM by aggregating influences from adjacent REM stages, specifically referencing two subsequent and one preceding REM epoch.

which correctly adjusts the predictions by considering the relationships of neighboring epochs.

An example of this can be seen in Fig. 4c, where the 321<sup>st</sup> epoch, initially misclassified as N2 due to displaying typical N2 characteristics in the heatmap, is correctly reclassified as REM when the *Inter-epoch ViT* aggregates influences from adjacent REM stages. Except for the first window where the target epoch is at the very end, the target epoch is accurately predicted as REM in all other windows where it is followed by REM stages. Upon analyzing the sliding window results and investigating the relevance scores between the target epoch and its neighboring epochs, it is evident that the two subsequent REM epochs exhibit the highest relevance scores, followed by a high relevance score for one preceding REM epoch. This pattern indicates that these two subsequent and one preceding REM epochs significantly contribute to the accurate correction of the target epoch's classification. For additional examples from other patients, see Supplementary Figs. 9–11.

We conducted a McNemar test to assess the statistical significance of the performance differences between the *Intra-epoch ViT* and *Inter-epoch ViT*, as reported in Table 3. This test utilizes the counts of discrepancies where the *Intra-epoch ViT* was correct but the *Inter-epoch ViT* was wrong (Intra-correct Inter-wrong) and vice versa (Intra-wrong Inter-correct) to compute the McNemar statistic. The chi-square ( $\chi^2$ ) values derived from this

test assess whether the observed differences in classification performance between the two settings are due to chance or are statistically significant. For instance, in the Wake stage, the *Intra-epoch ViT* correctly classifies 1,207 samples while the *Inter-epoch ViT* misclassifies them, and conversely, the *Inter-epoch ViT* correctly classifies 4,521 samples which the *Intra-epoch ViT* misclassifies. The resulting  $\chi^2$  of 1917.35 and a  $p$ -value of less than 0.05 (0.00E+00) indicate a statistically significant difference in performance, suggesting that the *Inter-epoch ViT* outperforms the *Intra-epoch ViT* at the Wake stage. Across all stages except N3, significant performance improvements are noted with the *Inter-epoch ViT*, particularly in the REM stage where the largest  $\chi^2$  highlights substantial gains from the *Inter-epoch ViT* assistance.

### Consistency evaluation of explainability

Until now, we have explored the explainability of *SleepXViT* through post-hoc explanation methods from explainable artificial intelligence (XAI) to elucidate and interpret the decisions made by our model. However, evaluating these XAI methods presents substantial challenges, as quantifying explanations or interpretability is complex and lacks standardized approaches. Although human expert evaluations are often utilized, their subjective nature can lead to misunderstandings, particularly when model's logic may be opaque to human analysts. To mitigate this, we assessed *SleepXViT*'s



Table 3 | McNemar Test Results

Stage	Intra-correct Inter-wrong	Intra-wrong Inter-correct	chi-square ( $\chi^2$ )	p-value
Wake	1207	4521	1917.35	0.00E+00
N1	4417	6887	539.71	0.00E+00
N2	7996	8506	15.76	7.18E-05
N3	2263	2154	2.69	1.01E-01
REM	745	4054	2281.62	0.00E+00
Overall	16628	26122	31542.74	0.00E+00

Evaluation of the statistical significance of performance differences between the *Intra-epoch ViT* and *Inter-epoch ViT*. The chi-square ( $\chi^2$ ) values derived from this test assess whether the differences in classification performance are statistically significant. A *p*-value less than 0.05 is considered statistically significant.

explainability by examining the consistency of its explanations across similar inputs, following the principle that similar inputs should yield similar explanations<sup>50</sup>. This analysis focuses on determining whether the heatmaps, which reveal the model’s reasoning, demonstrate consistent patterns when predictions belong to the same class.

We conducted *K*-means clustering on the heatmaps, setting *K* to five to match the number of sleep stages. The results, presented in Fig. 5a, indicate that certain clusters are predominantly characterized by specific classes, enabling us to assign pseudo labels based on the most common class within each cluster: Wake to Cluster #1, N2 to Cluster #3, N3 to Cluster #4, and REM to Cluster #5. Cluster #2 is an exception, displaying a relatively uniform distribution across all classes except REM, with a slightly higher occurrence of Wake over N1. Since Cluster #1 is predominantly categorized as Wake and no other cluster predominantly represents N1, we opted to assign a pseudo label of N1 to Cluster #2. Figure 5b shows randomly sampled heatmaps from each cluster, confirming that similar heatmaps tend to cluster together. This pattern holds for all classes except N1, where the

heatmap characteristics closely align with those depicted in the class-specific patterns shown in Fig. 3.

To strengthen our evaluation, we checked whether grouping heatmaps based on ground-truth labels showed consistent patterns. Supplementary Fig. 12 presents these ground-truth-based clusters with random samples for each class. Notably, the heatmaps in each ground-truth class resemble the clustering results of corresponding pseudo labels in Fig. 5b and Supplementary Fig. 13, confirming the reliability of the heatmaps generated by *SleepXViT*. For N1, the most heterogeneous cluster in our unsupervised clustering, the ground-truth-based cluster also exhibits the greatest diversity in heatmaps. This supports our observation that *SleepXViT* tend to predict N1 when no distinct class-specific patterns dominate. Finally, Fig. 5c illustrates the alignment between the model’s classification predictions and the clustering outcomes, highlighting significant accuracy in clusters characterized by distinct classes, such as N2 and REM. However, there is noticeable confusion between Wake and N1 clusters. Similarly, N3, which primarily analyzes the EEG area similar to Wake and N1, also shows slightly reduced accuracy.

Discussion

*SleepXViT* has demonstrated superior performance across various datasets compared to other baseline models, as shown in Table 1. In the case of *SleepTransformer*, although it employs a Transformer architecture similar to *SleepXViT*, there is a notable difference in performance. This difference can be attributed to the amount and quality of information contained in the input data, rather than the analysis being visual or non-visual. Unlike *SleepTransformer*, which relies solely on single-channel EEG, *SleepXViT* consolidates all available multimodal PSG signals into a standardized image format. This disparity in information density helps explain why the performance gap between the two models is even larger in the KISS dataset, which includes a greater number of multimodal PSG signals compared to the SHHS dataset. Notably, in the KISS dataset, *SleepTransformer* even underperforms relative to the simple CNN-based model developed by Jeong

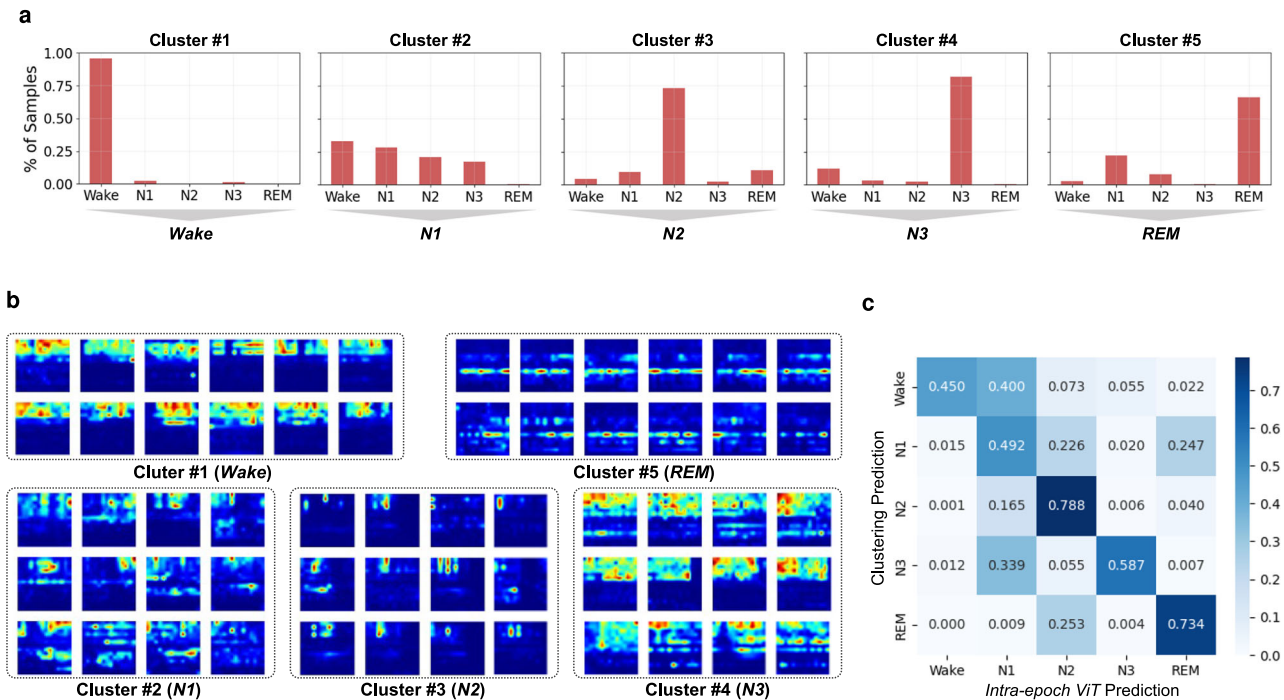
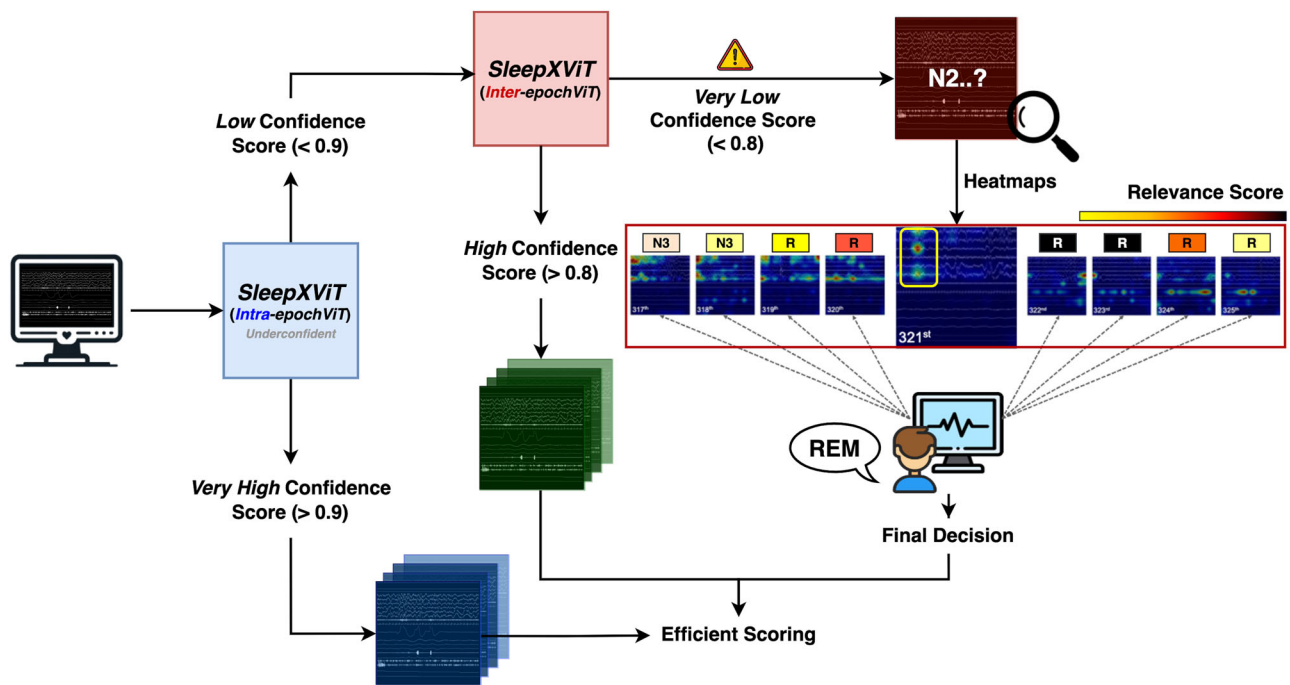


Fig. 5 | Evaluation of Explainability via Clustering. Heatmaps are clustered into five groups corresponding to the five sleep stages using *K*-means clustering with *K* = 5. **a** Distribution of predicted classes across clusters and results of cluster pseudo labeling. **b** Randomly sampled heatmaps from each cluster showing that similar maps cluster

together, supporting the model’s consistent explainability. The displayed sleep stage is a pseudo labeled sleep stage. **c** Normalized confusion matrix comparing the predictions of the *Intra-epoch ViT* with the clustering outcomes, illustrating the alignment and discrepancies between model predictions and clustering.





**Fig. 6 | Clinical Scenario of SleepXViT.** Using the confidence scores generated by the model, we can first filter images with *Intra-epoch ViT* identifying cases with low confidence for further review. These selected samples are then examined by the *Inter-epoch ViT* for additional filtering. Finally, images requiring human assessment

are flagged, allowing clinicians to make the final determination with the aid of heatmaps and relevance scores. With this workflow based on confidence, clinicians make the final decision and finish the staging task efficiently.

et al., which was trained on multimodal signal data, further underscoring that the information gap can outweigh the architectural advantages of a transformer.

According to Jeong et al., while both image-based and signal-based models utilize CNN and Bi-LSTM in their architecture, performance differences persist, with image-based models exhibiting greater robustness to variations in data collection environments. This underscores the complexity of developing a model that explains and comprehensively analyzes various types of raw time-series data, which may lead to suboptimal performance. In contrast, a visual approach simplifies model design by converting diverse types of signals into a single image format. Furthermore, when using the same image data, *SleepXViT* outperforms the CNN-based model by Jeong et al., likely due to the vision transformer's attention mechanism, which effectively captures two-dimensional relationships across both channel and time domains. This enables more efficient and comprehensive information learning compared to CNNs, which are better suited for capturing local features.

The KISS dataset provides a robust foundation with a large sample size and diverse sleep patterns. However, it primarily consists of male patients, reflecting a common demographic characteristic seen in sleep disorder clinics. This could introduce bias in representing other populations, such as females and the elderly. To address these concerns, we conducted an analysis, the results of which were generally favorable. Supplementary Note 3 and Supplementary Fig. 14 analyze the demographic bias of the dataset, revealing that age, BMI, and sex lead to statistical differences in performance. However, these differences are relatively small and do not significantly affect the overall distribution of performance across demographic groups, suggesting that *SleepXViT* has not learned in a highly biased manner. In addition, we analyzed outcomes for individuals with sleep-related disorders, including severe sleep fragmentation and OSA. Supplementary Fig. 15 statistically demonstrates that while the *SleepXViT*'s performance shows a slight decline in these cases, the impact is minimal and does not undermine the model's overall reliability.

As shown in Fig. 2, *SleepXViT* is almost perfectly calibrated, indicating that the model's confidence is linearly related to its accuracy (Fig. 2a). This

allows confidence to serve as a method for explaining the model's results and assessing the reliability of its predictions, enhancing practical usability. Thus, while *SleepXViT* should not be relied upon exclusively, it can effectively function as a reliable auxiliary tool. Figure 6 illustrates how *SleepXViT* can be practically utilized in clinical settings. An epoch's multimodal PSG data, converted into a standardized image, first passes through the *Intra-epoch ViT*. As shown in Fig. 2a, the *Intra-epoch ViT* exhibits a slight tendency towards underconfidence compared to the *Inter-epoch ViT*, indicating that its high-confidence (e.g., > 0.9) predictions are likely to be reliable. These confidently predicted epochs can bypass further processing, reducing inference time without sacrificing accuracy. Notably, the *Intra-epoch ViT* shows comparable performance to the *Inter-epoch ViT* for certain classes, such as N3 and Wake. If the *Intra-epoch ViT*'s confidence level is not sufficiently high, the epoch undergoes further processing by the *Inter-epoch ViT*. If the *Inter-epoch ViT* also flags the epoch as ambiguous (e.g., confidence score below 0.8), clinicians can focus on the potentially unclear case. For flagged epochs, clinicians can refer to *SleepXViT*'s heatmaps and relevance scores from both the *Intra*- and *Inter*-ViTs, enabling them to make informed decisions about whether to accept or adjust the *SleepXViT*'s predictions.

*SleepXViT* utilizes PSG images similar to the data observed by human experts during 'visual scoring', generates results, and provides visual explanations through heatmaps to enhance intuitive understanding. We have assessed whether these explanations align with AASM<sup>11,12</sup> rules, the standard criteria for human scoring. According to AASM guidelines, only EEG, EOG, and EMG channels are relevant for sleep staging. However, we have included ECG and respiratory-related signals in hopes of finding scientific clues about their effect on sleep staging.

In the Wake stage, beta rhythms (>13 Hz) predominantly appear in the frontal EEG areas, and alpha rhythms (8–12 Hz) dominate in the occipital areas. Frequent blinking and combined eye movements (0.5–2 Hz) are observed. The heatmap for the Wake in Fig. 3a shows all EEG regions highlighted, indicating a comprehensive assessment of beta and alpha waves in the central and occipital regions. The EOG's relevance score is also slightly elevated, reflecting eye movements. The N1 stage serves as a transitional

phase from wakefulness to deeper sleep stages, characterized by the disappearance of alpha rhythms and the emergence of slow, conjugate roving eye movements. The EEG displays medium amplitude and mixed frequencies, predominantly showing theta waves (4–7 Hz), with irregularly spaced slow waves. Breathing generally becomes slower and deeper than during wakefulness. As shown in Fig. 3a, the relevance is higher in the central regions compared to the occipital regions, where alpha is more prominent. EOG is highlighted, and the Flow area related to respiration is notably emphasized. This stage is often confused with Wake or N2, and recent studies indicate that  $45.2 \pm 9.2\%$  of epochs scored as N1 by one technologist did not receive agreement from other scorers, with only  $1.5 \pm 1.1\%$  scored unanimously<sup>51</sup>. This high level of inter-scorer disagreement suggests that our model's difficulty with N1 classification aligns with broader scoring challenges. As mentioned in the Results section, N2 shows characterized by prominent features in the EEG such as sleep spindles and K-complexes, and *SleepXViT*'s heatmap effectively reflects this (Fig. 3b). In summary, the heatmaps for Wake, N1, and N2 can be sufficiently interpreted using the AASM rules for sleep staging.

Meanwhile, the N3 and REM stages require deeper analysis as they are not fully interpreted by AASM guidelines. N3, representing the deepest sleep stage, is characterized by high amplitude delta slowing within the range of 0.5–2 Hz, predominantly measured in the frontal to central regions. However, the heatmaps displayed in Fig. 3 reveal that all EEG regions are strongly highlighted, and notably, the relevancy of the EOG signals is also elevated. This is unexpected since EOG is not typically associated with N3, where no eye movements should occur. This phenomenon could be due to the proximity of EOG and EEG sensors causing high-voltage slow delta waves to leak and interfere<sup>52,53</sup>. *SleepXViT* appears to focus on the EOG area to determine whether the signals are due to actual eye movements or EEG interference. The REM stage is characterized by rapid eye movements and suppressed chin EMG activity, along with low-amplitude mixed frequency EEG that appears without K-complexes or sleep spindles. Therefore, it was expected that the model would primarily scrutinize the EOG and EMG patterns when making predictions. However, *SleepXViT* tends to prioritize the Flow over the EOG and EMG areas. During REM, it is known that both the rate and variability of respiration increase compared to other stages, and irregular breathing patterns are associated with rapid eye movements<sup>54,55</sup>. It seems that *SleepXViT* analyzes these subtle respiratory patterns, which are difficult to detect through human observation, and uses them to determine the REM stage. Upon reviewing samples classified as REM, it was evident that considerable attention was paid to the Flow throughout the 30-second epoch, likely in an attempt to evaluate these irregularities.

We investigated whether *SleepXViT* provides consistent visual explanations across similar data by performing clustering analysis, which confirmed significant class-specific clustering. This indicates that the model explains its predictions based on fairly consistent criteria and logic. However, the cluster pseudo-labeled as N1 showed an even distribution across all classes except REM (Fig. 5a), suggesting that the N1 heatmaps might contribute to confusion with other stages. Indeed, considering the presence of N1 in clusters primarily assigned to REM, it appears that N1's heatmaps are indistinguishable from other stages. The alignment between prediction results and clustering outcomes presented in Fig. 5c is generally lower than the actual class-wise accuracy, highlighting the limitations of this evaluation method. For instance, high relevance scores in certain EEG areas could indicate different phenomena based on the class: identifying alpha and beta waves in Wake, transitioning to theta waves in N1, detecting sleep spindles or K-complexes in N2, observing delta waves in N3, or differentiating mixed frequency brain waves in REM which similar to N1. These findings underscore the need for more refined evaluation methods to accurately assess the model's internal consistency.

To explore the clinical value of these heatmaps, we analyzed them alongside confidence scores to investigate how heatmap patterns might indicate prediction reliability. In Supplementary Fig. 16, we aggregated heatmaps of correct predictions by 0.1 confidence intervals for each class, with 1,000 heatmaps per interval, focusing on confidence scores above 0.4

due to limited samples at lower scores. For Wake, N1, and N2 stages, we observed that as confidence scores increase, the intensity of captured visual cues within the same signals also increases. In contrast, for REM, a more diverse trend was observed: Flow intensity increases with confidence scores, while EMG remains relatively consistent, except in the highest confidence range (0.9–1.0). EOG is substantially attended only within the highest confidence range. In N3, both EEG and EOG are highly attended across all confidence scores. These findings suggest that clinicians might estimate prediction accuracy based on the intensity of visual cues or the types of signals highlighted in the heatmaps. For instance, if *SleepXViT* predicts REM without EOG activity in the heatmap, a clinician could infer a prediction accuracy below 90%. Supplementary Fig. 17 displays aggregated heatmaps from model predictions and ground truth labels. Incorrect predictions typically show weaker signal intensities in the heatmaps, illustrating that heatmaps alone cannot fully determine prediction accuracy. For instance, heatmaps for REM predictions look similar regardless of their correctness. However, combining heatmaps with confidence scores can enhance the assessment of model reliability, allowing clinicians to better understand and trust the model's low-confidence predictions in clinical applications.

Human scorers perform sleep staging by considering continuity and patterns across multiple epochs, as a single epoch alone often fails to capture the entire context of sleep stages. Similarly, we have emulated this approach in *SleepXViT* by integrating the *Inter-epoch ViT*, which has proven crucial in enhancing *SleepXViT*'s performance. To determine an appropriate sequence length, we conducted experiments by varying the input sequence length, as shown in Table 2. However, we did not observe a distinct trend regarding performance changes. Consequently, we selected 10 epochs, which yielded the best overall Macro F1 score. In *SleepXViT*, the *Inter-epoch ViT* utilizes a sliding-window ensemble technique during inference. With a sequence length of 10 epochs, the prediction for the target epoch incorporates information from the 9 preceding and 9 succeeding epochs, resulting in a total of 19 epochs considered for each prediction, including the target epoch. This approach is consistent with finding from other models; for example, SleepTransformer<sup>18</sup> empirically used an input sequence length of 21, while other models such as DeepSleepNet<sup>56</sup> and SeqSleepNet<sup>57</sup> selected a sequence length of 20. This indicates that a sequence length of around 20 epochs is generally deemed appropriate in several studies, indirectly supporting the rationale for selecting 10 epochs in our model.

As shown in the McNemar Test results in Table 3, *Inter-epoch ViT* plays a particularly significant role in predicting REM stages. According to AASM rules, once the REM stage is scored, it continues to be scored as REM until there are clear criteria to stop, even if rapid eye movements are not present. Therefore, even if a single epoch does not exhibit distinct REM characteristics, it is scored as REM if adjacent epochs are clearly REM. Figure 4b and c demonstrate that *SleepXViT* follows this rule precisely. When using the *Inter-epoch ViT*, the target epoch, initially predicted by *Intra-epoch ViT* as N2, is correctly classified as REM when considering adjacent epochs. Particularly, when the target epoch is at the end of a multi-epoch sequence, it remains predicted as N2, but it is correctly adjusted to REM if followed by clear REM epochs. This correction occurs because, although rapid eye movements may not be distinctly observed, the continuation of REM stages before and after the target epoch leads to its prediction as REM. The relevance scores of surrounding epochs show that the two subsequent REM epochs and the one preceding REM epoch play a crucial role in maintaining the continuity of REM scoring. Additionally, as seen in Table 3, the *Inter-epoch ViT* significantly influences the detection of Wake stages. Wake is particularly often confused with N1 due to the ambiguous transitions between these stages; thus, examining adjacent epochs aids in clarification. Indeed, the implementation of the *Inter-epoch ViT* in *SleepXViT* has yielded significant accuracy improvements in detecting Wake, N1, and REM stages (Fig. 2d).

In theory, multimodal time-series PSG data in its raw form captures detailed information, and developing a model specifically tailored for these heterogeneous data types could lead to the most accurate sleep staging.

However, visual representation, although potentially containing slightly less granular information than raw signals, offer significant practical advantages in our particular clinical application:

- **Information Quality:** Clinicians use visual cues to label sleep stages, which serve as ground truth for this task. Therefore, visual representations must contain enough information to replicate the clinicians' sleep staging process; any additional information in raw time-series signals is likely redundant.
- **Consistent Preprocessing Across Different Systems:** We use a standardized preprocessing pipeline to convert time-series data into images, avoiding the need to adjust hyperparameters for each system due to variations in signal sampling rates and scales.
- **Simplifying Design through Modality Reduction:** Converting various signals into a unified image format greatly simplifies model design, enabling to easily design an attention mechanism that effectively captures relationships across both channel and time domains.
- **Leveraging Advances in Computer Vision:** Analyzing visual representations opens the gate to leverage the latest techniques in computer vision, one of the fastest-evolving fields of AI, for the sleep staging task. This bridging can accelerate innovation in this area.
- **Explainability Aligned with Clinical Practice:** By learning the 2-dimensional relationships in PSG images, *SleepXViT*'s explanations align more closely with clinicians' perspectives (AASM guidelines), providing interpretable insights across all relevant channels.

We believe that these benefits of visual representations are generally applicable to other tasks where human experts label time-series data based on visual cues.

## Methods

### Data preprocessing

The KISS dataset follows the AASM standard for sleep staging with 5 classes (Wake, N1, N2, N3, REM), while the SHHS dataset adheres to the R&K standard for sleep staging. Therefore, to score the datasets consistently, the S3 and S4 stages in the SHHS dataset were combined and labeled as N3 stage. Additionally, epochs labeled as 'Movement' or 'Unknown' were removed during preprocessing.

The KISS dataset was obtained from four medical centers. Ethical approval was granted by the institutional review boards of Hallym University College of Medicine (Chuncheon, Republic of Korea: approval number 2020-03-022), Seoul National University Bundang Hospital, (Seongnam, Republic of Korea: approval number C-2007-179-1143), and Seoul National University Hospital (Seoul, Republic of Korea: approval number B-2010/640-401) approved this study. As the dataset was de-identified, written informed consent was not required<sup>32</sup>.

The KISS and SHHS datasets were converted from EDF format to images using a Node.js-based conversion module. The signals were undergone a *standardized* preprocessing pipeline as outlined in Jeong et al.<sup>32</sup>. This process involves several important steps:

1. **Signal Order Adjustment:** To maintain a standard format, the signals are consistently ordered across all images. If a particular signal is absent, the corresponding area is left blank.
2. **Filtering:** Each signal is filtered according to its specific frequency range, capturing only relevant information. For the KISS sampled at 200 Hz, we used a High Pass/Low Pass/Band Stop filter with the following frequencies: EOG: 0.3/35/60Hz, EMG: 10/70/60Hz, ECG: 0.3/70/60Hz<sup>58</sup>. If the sampling frequency was lower than 200Hz, the cutoff frequency was set to half of the sampling frequency.
3. **Unit Normalization:** All signal units are standardized for consistency.
4. **Min-Max Scaling:** All signals undergo min-max scaling to ensure uniformity in visual data representation.

This standardized preprocessing minimized variability in image inputs due to software or display settings. In addition, the same preprocessing was applied to SHHS, demonstrating its robustness as standardization approach.

After converting signal data into image format, we resized and cropped all images to  $224 \times 224$  with only 14 channels: EEG (C3-M2, C4-M1, O1-M2, O2-M1), and EOG (E1-M2, E2-M1), Chin EMG, ECG, Flow, Thermister, Thoracic Movement, Abdominal Movement, and Oxygen Saturation (85 ~100% and 40~100%). Except for the Flow and two Oxygen Saturation channels, the rest of 11 channels have a height of 43 pixels out of the original image's height of 1080 pixels. To make all channels have the same height, the Flow signal, which occupies 129 pixels (3 channels' height), and the two oxygen saturation signals, each occupying 86 pixels (2 channels' height), are resized to 43 pixels per channel. Subsequently, the signals are stacked in order. Finally, the resulting  $602 \times 1920$  image, with all 14 channels having the same height, is resized to a size of  $224 \times 224$ . *SleepXViT* utilizes input images by dividing them into  $16 \times 16$  patches. This preprocessing ensures that each patch contains only a portion of information from a single channel, rather than mixing multiple channel data. Examples of the PSG image format and preprocessing results are shown in Supplementary Fig. 18.

### SleepXViT Training

As illustrated in Fig. 1, *SleepXViT* comprises the *Intra-epoch ViT*, which learns features from the input PSG images, and the *Inter-epoch ViT*, which learns the relationships between adjacent epochs and predicts the sleep stage sequences for multiple epochs. The implementation details and training processes of each module are described below, and the related figures can be found in Fig. 7.

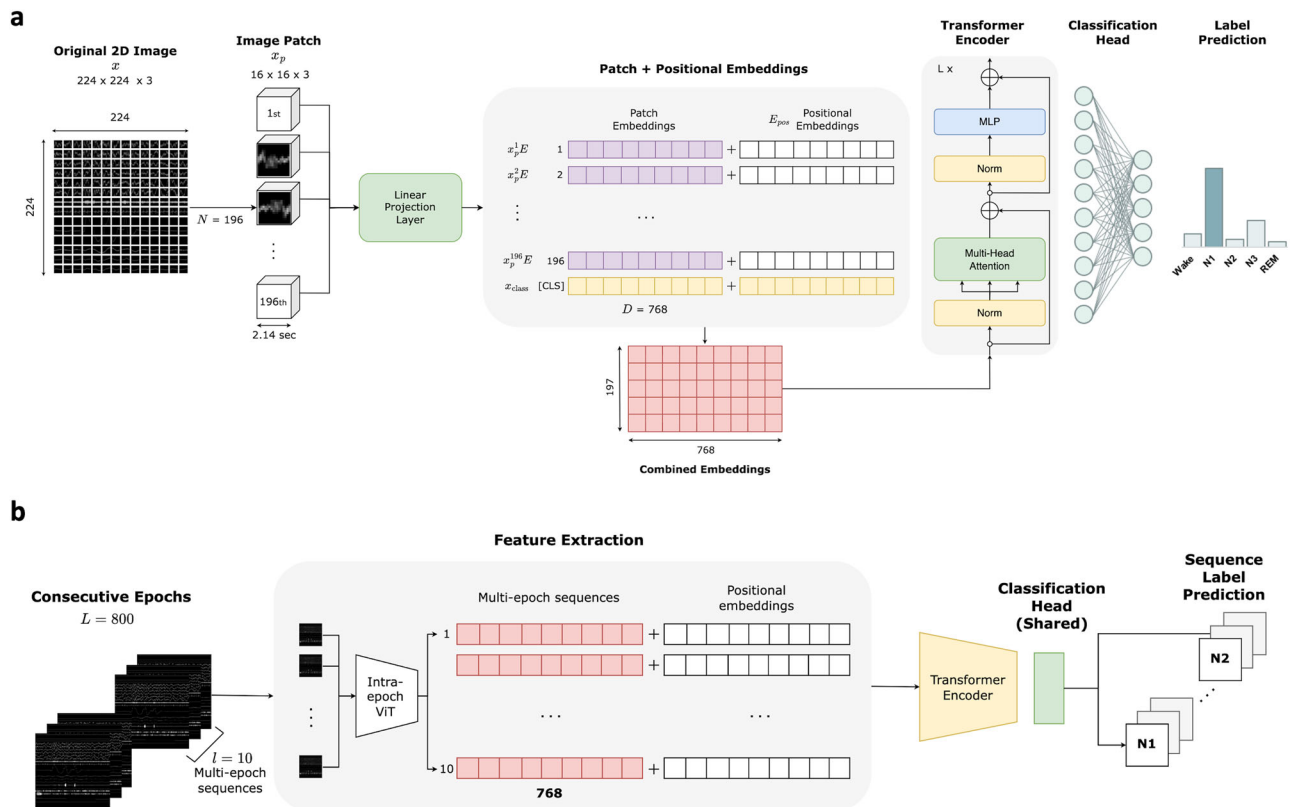
In Fig. 7a, the *Intra-epoch ViT*, which utilizes the Vision Transformer (ViT) architecture<sup>34</sup>, is illustrated. It begins by processing the 2D epoch image  $x \in \mathbb{R}^{H \times W \times C}$  by reshaping it into a sequence of flattened 2D patches  $x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$ . Here,  $(H, W)$  represents the resolution of the image,  $C$  is the number of channels,  $P^2$  specifies the resolution of each image patch, and  $N = \frac{H \times W}{P^2}$  is the total number of patches. This transformation adapts the input appropriately for the transformer encoder. To effectively utilize the ViT architecture and leverage its pre-trained weights, we opted for a patch size of  $16 \times 16$ . Standardized PSG images are segmented into patches according to the ViT methodology, ensuring that each patch contains information from a single channel. Each patch is of size  $16 \times 16 \times 3$ , resulting in a dimension  $D$  of 768. Given our input image size of  $224 \times 224 \times 3$ , the total number of patches  $N$  is calculated as 196, which results from the formula  $N = \frac{224 \times 224}{16 \times 16}$ . These patches cover 2.14 seconds of waveform from one channel. For classification purposes, an additional learnable classification token [CLS] is included in the sequence. Positional embeddings are added to the patch embeddings, resulting in combined embeddings size of  $E_{pos} \in \mathbb{R}^{(N+1) \times D}$ . The resulting embeddings is then processed by a standard Transformer encoder. Having confirmed the superior performance of the pre-trained ViT over training from scratch, we utilize an off-the-shelf pre-trained ViT. This model employs weights from the official Google JAX implementation, pre-trained on ImageNet-21k and fine-tuned on ImageNet-1k.

In Fig. 7b, the second stage of the training process is depicted. Feature embeddings  $z \in \mathbb{R}^D$  for each epoch are extracted using the *Intra-epoch ViT* encoder, and these embeddings are compiled into sequences. Assuming there are  $L$  epochs (e.g. 800) of PSG images from one night, we can create  $M = \frac{L}{l}$  (e.g. 80) multi-epoch sequences, each consisting of consecutive  $l = 10$  epochs. The *Inter-epoch ViT* utilizes these multi-epoch sequences by adding learnable positional embeddings for training. We applied multi-head attention in the sequence dimension using the *Inter-epoch ViT* encoder, which produced a vector of the same size as the input. The *Inter-epoch ViT* classifier is then trained to make predictions for each epoch in the sequence, aiming to minimize the cross-entropy loss across  $l$  consecutive epochs:

$$L_{inter} = -\frac{1}{l} \sum_{i=1}^l y_i \log(\hat{y}_i) \quad (1)$$

Details of hyperparameters and optimization methods are provided in Supplementary Table 1.





**Fig. 7 | SleepXViT training process.** The SleepXViT training process consists of two stages: *Intra-epoch ViT* and *Inter-epoch ViT*. **a** The *Intra-epoch ViT* training, which employed the Vision Transformer (ViT) architecture<sup>34</sup> to learn the embedding of the epoch image. **b** The *inter-epoch ViT* training, uses the frozen *Intra-epoch ViT* to

generate embedded vectors. These vectors are then input into another transformer encoder, which generates the consecutive  $l$  epochs from referencing the adjacent epochs.

### Ensemble with Sliding Window

During training, we employed sequences of  $l$  non-overlapping epochs and observed potential biases in the initial and final tokens within these sequences. These biases arise because these tokens are influenced solely by either preceding or succeeding epochs. To mitigate this issue, we utilized a sliding-window approach during inference, using a stride of one epoch, as depicted in Fig. 4a. This adaptation ensures each epoch is evaluated within a comprehensive context, considering  $l$  adjacent epochs around the target epoch (epoch of interest). As a result, the model generates/softmax values for each epoch prediction. By aggregating these values across all epochs, we enhanced the model's accuracy, improving the Macro F1 score by an impressive 1%. This sliding-window ensemble technique not only bolsters performance but also provides deeper insights into the decision-making process of the *Inter-epoch ViT*, resulting in a more robust and context-aware framework for sleep stage classification.

### McNemar Test

The McNemar test is a statistical method used for analyzing categorical data for two related samples<sup>59</sup>. It is primarily applied to assess the before-and-after changes in the same subjects across two treatments, conditions, techniques, or periods. This test checks the independence of two related sample data and focuses specifically on the differences between two binary variables. We conducted this test to observe the differences when the *Inter-epoch ViT* was applied versus when it was not. The statistic for the McNemar test is calculated as follows:

$$Q = \frac{(b - c)^2}{b + c} \quad (2)$$

This statistic follows a chi-squared ( $\chi^2$ ) distribution with one degree of freedom. The McNemar test assesses whether the differences between 'b' and 'c' are statistically significant; a significant difference indicates a statistically significant difference between the two conditions. In our application of the McNemar test, 'b' represents the cases where the *Intra-epoch ViT* was correct but the *Inter-epoch ViT* was wrong, and 'c' represents the cases where the *Intra-epoch ViT* was wrong but the *Inter-epoch ViT* was correct.

### Clustering of heatmaps

The entire KISS test dataset was used for the heatmap clustering analysis. Due to hardware memory limitations, it was challenging to perform Principal Component Analysis (PCA) on the full dataset simultaneously. To alleviate the computational burden, we randomly sampled heatmap images to preserve the original label distribution across different sleep stages, allocating 45 images for Wake, 30 for N1, 70 for N2, 25 for N3, and 30 for REM per patient. Additionally, we resized the images from  $224 \times 224 \times 3$  to  $32 \times 32 \times 3$ . Subsequently, we applied PCA to extract 100 salient features from the entire dataset, which were then clustered into five distinct groups using K-means clustering ( $K=5$ ).

### Data availability

The full dataset of KISS does not publicly available because of legal restrictions imposed by Korean government in relation to the Person Information Protection Act. However, if some investigators wish to use it, they could access it after obtaining the relevant permit from the Korean National Information Society Agency.

### Code availability

The developed code used for training and visualization of SleepXViT is available at <https://github.com/sarahhyojin/SleepXViT>.

Received: 18 September 2024; Accepted: 10 December 2024;  
Published online: 25 January 2025

## References

- Benington, J. H. & Heller, H. C. Restoration of brain energy metabolism as the function of sleep. *Prog. Neurobiol.* **45**, 347–360 (1995).
- Moldofsky, H. Sleep and the immune system. *Int. J. Immunopharmacol.* **17**, 649–654 (1995).
- Walker, M. P. The role of sleep in cognition and emotion. *Ann. N. Y. Acad. Sci.* **1156**, 168–197 (2009).
- Astill, R. G., Van der Heijden, K. B., Van IJzendoorn, M. H. & Van Someren, E. J. Sleep, cognition, and behavioral problems in school-age children: a century of research meta-analyzed. *Psychol. Bull.* **138**, 1109 (2012).
- Cappuccio, F. P., Cooper, D., D'Elia, L., Strazzullo, P. & Miller, M. A. Sleep duration predicts cardiovascular outcomes: a systematic review and meta-analysis of prospective studies. *Eur. Heart J.* **32**, 1484–1492 (2011).
- Cappuccio, F. P., D'Elia, L., Strazzullo, P. & Miller, M. A. Quantity and quality of sleep and incidence of type 2 diabetes: a systematic review and meta-analysis. *Diab. Care* **33**, 414–420 (2010).
- Cappuccio, F. P. et al. Meta-analysis of short sleep duration and obesity in children and adults. *Sleep* **31**, 619–626 (2008).
- Weitzman, E. D., Kripke, D. F., Goldmacher, D., McGregor, P. & Nogueira, C. Acute reversal of the sleep-waking cycle in man: Effect on sleep stage patterns. *Arch. Neurol.* **22**, 483–489 (1970).
- Yoon, D. W. & Shin, H.-W. Sleep tests in the non-contact era of the covid-19 pandemic: Home sleep tests versus in-laboratory polysomnography. *Clin. Exp. Otorhinolaryngol.* **13**, 318–319 (2020).
- Rundo, J. V. & Downey, R. Chapter 25 - polysomnography. In Levin, K. H. & Chauvel, P. (eds.) *Clinical Neurophysiology: Basis and Technical Aspects*, vol. 160 of *Handbook of Clinical Neurology*, 381–392 (Elsevier, 2019). <https://www.sciencedirect.com/science/article/pii/B9780444640321000254>.
- Berry, R. B. et al. The aasm manual for the scoring of sleep and associated events. *Rules Terminol. Tech. Specif. Darien, Ill., Am. Acad. Sleep. Med.* **176**, 7 (2012).
- Berry, R. B. et al. Aasm scoring manual updates for 2017 (version 2.4) (2017).
- Patel, A. K., Reddy, V., Shumway, K. R. & Araujo, J. F. Physiology, sleep stages. In *StatPearls [Internet]* (StatPearls Publishing, 2022).
- Younes, M., Raneri, J. & Hanly, P. Staging sleep in polysomnograms: analysis of inter-scorer variability. *J. Clin. Sleep. Med.* **12**, 885–894 (2016).
- Anderer, P. et al. An e-health solution for automatic sleep classification according to Rechtschaffen and kales: validation study of the somnolyzer 24 × 7 utilizing the siesta database. *Neuropsychobiology* **51**, 115–133 (2005).
- Berthomier, C. et al. Automatic analysis of single-channel sleep eeg: validation in healthy individuals. *Sleep* **30**, 1587–1595 (2007).
- Chapotot, F. & Becq, G. Automated sleep-wake staging combining robust feature extraction, artificial neural network classification, and flexible decision rules. *Int. J. Adapt. Control Signal Process.* **24**, 409–423 (2010).
- Phan, H. et al. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Trans. Biomed. Eng.* **69**, 2456–2467 (2022).
- Phan, H. et al. L-seqsleepnet: Whole-cycle long sequence modelling for automatic sleep staging. *IEEE J. Biomed. Health Inform.* **27**, 4748–4757 (2023).
- Chen, Z. et al. Automated sleep staging via parallel frequency-cut attention. *IEEE Trans. Neural Syst. Rehabil. Eng.* **31**, 1974–1985 (2023).
- Eldele, E. et al. An attention-based deep learning approach for sleep stage classification with single-channel eeg. *IEEE Trans. Neural Syst. Rehabil. Eng.* **29**, 809–818 (2021).
- Phan, H. et al. Xsleepnet: Multi-view sequential model for automatic sleep staging. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 5903–5915 (2021).
- Dai, Y. et al. Multichannelsleepnet: A transformer-based model for automatic sleep stage classification with psg. *IEEE J. Biomed. Health Inform.* **27**, 4204–4215 (2023).
- Amann, J. et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* **20**, 1–9 (2020).
- Jia, X., Ren, L. & Cai, J. Clinical implementation of ai technologies will require interpretable ai models. *Med. Phys.* 1–4 (2020).
- Muto, V. & Berthomier, C. Looking for a balance between visual and automatic sleep scoring. *npj Digital Med.* **6**, 165 (2023).
- Danker-Hopfe, H. et al. Interrater reliability between scorers from eight european sleep laboratories in subjects with different sleep disorders. *J. sleep. Res.* **13**, 63–69 (2004).
- Danker-Hopfe, H. et al. Interrater reliability for sleep scoring according to the rechtschaffen & kales and the new aasm standard. *J. sleep. Res.* **18**, 74–84 (2009).
- Guillot, A., Sauvet, F., During, E. H. & Thorey, V. Dream open datasets: Multi-scored sleep datasets to compare human and automated sleep staging. *IEEE Trans. neural Syst. Rehabil. Eng.* **28**, 1955–1965 (2020).
- Horie, K. et al. Automated sleep stage scoring employing a reasoning mechanism and evaluation of its explainability. *Sci. Rep.* **12**, 12799 (2022).
- Peng, L. et al. Sleepvitransformer: Patch-based sleep spectrogram transformer for automatic sleep staging. *Biomed. Signal Process. Control* **86**, 105203 (2023).
- Jeong, J. et al. Standardized image-based polysomnography database and deep learning algorithm for sleep-stage classification. *Sleep* **46**, zsad242 (2023).
- Bany Muhammad, M. & Yeasin, M. Eigen-cam: Visual explanations for deep convolutional neural networks. *SN Comput. Sci.* **2**, 47 (2021).
- Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR* (2021).
- Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30** (2017).
- AIHub. Image of sleep quality assessment and sleep disorder diagnosis. <https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=realm&dataSetSn=210> (2021). (Accessed on 11/01/2023).
- Ahn, H. K., Kang, Y. J., Yoon, W. & Shin, H.-W. Analysing the impact of body position shift on sleep architecture and stage transition: A comprehensive multidimensional study using event-synchronised polysomnography data. *J. Sleep Res.* (2023).
- Zhang, G.-Q. et al. The national sleep research resource: towards a sleep data commons. *J. Am. Med. Inf. Assoc.* **25**, 1351–1358 (2018).
- Quan, S. F. et al. The sleep heart health study: design, rationale, and methods. *Sleep* **20**, 1077–1085 (1997).
- Fiorillo, L., Favaro, P. & Faraci, F. D. Deepsleepnet-lite: A simplified automatic sleep stage scoring model with uncertainty estimates. *IEEE Trans. neural Syst. Rehabil. Eng.* **29**, 2076–2085 (2021).
- Van Gorp, H., van Gilst, M. M., Fonseca, P., Overeem, S. & van Sloun, R. J. Modeling the impact of inter-rater disagreement on sleep statistics using deep generative learning. *IEEE Journal of Biomedical and Health Informatics* (2023).
- Heremans, E. R. et al. U-pass: An uncertainty-guided deep learning pipeline for automated sleep staging. *Comput. Biol. Med.* **171**, 108205 (2024).
- Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, 1321–1330 (PMLR, 2017).

44. Mehrtash, A., Wells, W. M., Tempany, C. M., Abolmaesumi, P. & Kapur, T. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Trans. Med. Imaging* **39**, 3868–3878 (2020).
45. Van der Velden, B. H., Kuijff, H. J., Gilhuijs, K. G. & Viergever, M. A. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Med. Image Anal.* **79**, 102470 (2022).
46. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929 (2016).
47. Selvaraju, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626 (2017).
48. Montavon, G., Binder, A., Lapuschkin, S., Samek, W. & Müller, K.-R. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning* 193–209 (2019).
49. Chefer, H., Gur, S. & Wolf, L. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition*, 782–791 (2021).
50. Nauta, M. et al. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Comput. Surv.* **55**, 1–42 (2023).
51. Younes, M. et al. Reliability of the american academy of sleep medicine rules for assessing sleep depth in clinical practice. *J. Clin. Sleep. Med.* **14**, 205–213 (2018).
52. Estrada, E. et al. Itakura distance: A useful similarity measure between eeg and eog signals in computer-aided classification of sleep stages. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, 1189–1192 (IEEE, 2006).
53. Fan, J., Sun, C., Long, M., Chen, C. & Chen, W. Eognet: A novel deep learning model for sleep stage classification based on single-channel eog signal. *Front. Neurosci.* **15**, 573194 (2021).
54. Chung, G. S. et al. Rem sleep estimation only using respiratory dynamics. *Physiol. Meas.* **30**, 1327 (2009).
55. Orem, J. Neuronal mechanisms of respiration in rem sleep. *Sleep* **3**, 251–267 (1980).
56. Supratak, A., Dong, H., Wu, C. & Guo, Y. Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE Trans. neural Syst. Rehabil. Eng.* **25**, 1998–2008 (2017).
57. Phan, H., Andreotti, F., Cooray, N., Chén, O. Y. & De Vos, M. Seqsleepnet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Trans. Neural Syst. Rehabil. Eng.* **27**, 400–410 (2019).
58. Soh, J. H., Kang, Y. J., Yoon, W.-H., Park, C.-S. & Shin, H.-W. Positional obstructive sleep apnea and periodic limb movement during sleep: A large multi-center study. *Clin. Exp. Otorhinolaryngol.* (2024). <https://pubmed.ncbi.nlm.nih.gov/38693650/>.
59. Pembury Smith, M. Q. & Ruxton, G. D. Effective use of the mcnemar test. *Behav. Ecol. Sociobiol.* **74**, 1–9 (2020).

## Acknowledgements

This research was supported partly by the Creative-Pioneering Researchers Program through Seoul National University, partly by the Al-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-2023-00212780), and partly by the NRF of Korea

grant funded by Korea government (MSIT) (No. RS-2023-00222663). This research was 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.

## Author contributions

H.L. contributed to the research design, algorithm and model architecture development, implementation of explainability code, and conducted experiments on model training and visualization. J.J. developed a baseline model for performance comparison. H.K.L. contributed to dataset preparation and processing. J.J. and H.K.L. implemented visualization methods for the baseline model. J.H. proposed the evaluation method for explainability and conducted statistical analyses of heatmaps. Y.R.C. performed statistical analyses and interpreted results using confidence scores. Y.R.C., H.L., and H.K.L. led the manuscript writing. H.-W.S. and H.S.K. supervised the project, assisted with project design, and provided critical feedback on the manuscript.

## Competing interests

H.L. declares no competing interests. Y.R.C. declares no competing interests. H.K.L. declares no competing interests. J.J. declares no competing interests. J.H. declares no competing interests. H.S.K. declares no competing interests. 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.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41746-024-01378-0>.

**Correspondence** and requests for materials should be addressed to Hyun-Woo Shin or Hyung-Sin Kim.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025