SLEEP, 2024, 47, 1–11
https://doi.org/10.1093/sleep/zsae202
Advance access publication 31 August 2024
Original Article

Original Article

Refining sleep staging accuracy: transfer learning coupled with scorability models

Wolfgang Ganglberger[1,2,3,*,‡], Samaneh Nasiri[2,3,4,‡], Haoqi Sun[1,2,3], Soriul Kim[5], Chol Shin[5,6], M. Brandon Westover[1,2,3,‡] and Robert J. Thomas[3,7,*,‡]

[1]Department of Neurology, Beth Israel Deaconess Medical Center, Boston, MA, USA,
[2]McCance Center for Brain Health, Massachusetts General Hospital, Boston, MA, USA,
[3]Division of Sleep Medicine, Harvard Medical School, Boston, MA, USA,
[4]Biomedical Informatics & Neurology, Emory School of Medicine, Atlanta, GA, USA,
[5]Institute of Human Genomic Study, College of Medicine, Kore University, Seoul, Republic of Korea,
[6]Biomedical Research Center, Korea University Ansan Hospital, Ansan, Republic of Korea and
[7]Division of Pulmonary Critical Care & Sleep Medicine, Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA

[‡]These authors contributed equally.
[*]Corresponding authors. Wolfgang Ganglberger, Department of Neurology, Beth Israel Deaconess Medical Center, Boston, MA, USA. Email: wganglbe@bidmc.harvard.edu; Robert J Thomas, Division of Pulmonary Critical Care & Sleep Medicine, Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA. Email: rthomas1@bidmc.harvard.edu.

Abstract

Study Objectives: This study aimed to (1) improve sleep staging accuracy through transfer learning (TL), to achieve or exceed human inter-expert agreement and (2) introduce a scorability model to assess the quality and trustworthiness of automated sleep staging.

Methods: A deep neural network (base model) was trained on a large multi-site polysomnography (PSG) dataset from the United States. TL was used to calibrate the model to a reduced montage and limited samples from the Korean Genome and Epidemiology Study (KoGES) dataset. Model performance was compared to inter-expert reliability among three human experts. A scorability assessment was developed to predict the agreement between the model and human experts.

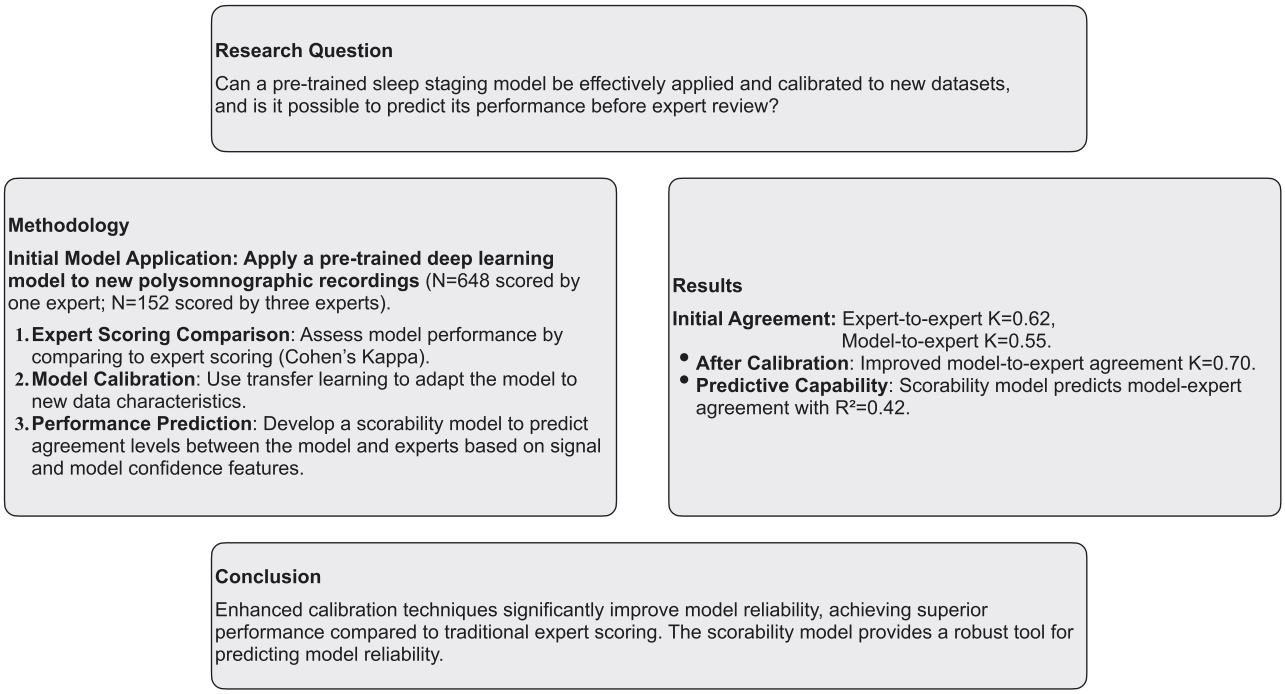
Results: Initial sleep staging by the base model showed lower agreement with experts ($\kappa = 0.55$) compared to the inter-expert agreement ($\kappa = 0.62$). Calibration with 324 randomly sampled training cases matched expert agreement levels. Further targeted sampling improved performance, with models exceeding inter-expert agreement ($\kappa = 0.70$). The scorability assessment, combining biosignal quality and model confidence features, predicted model-expert agreement moderately well ($R^2 = 0.42$). Recordings with higher scorability scores demonstrated greater model-expert agreement than inter-expert agreement. Even with lower scorability scores, model performance was comparable to inter-expert agreement.

Conclusions: Fine-tuning a pretrained neural network through targeted TL significantly enhances sleep staging performance for an atypical montage, achieving and surpassing human expert agreement levels. The introduction of a scorability assessment provides a robust measure of reliability, ensuring quality control and enhancing the practical application of the system before deployment. This approach marks an important advancement in automated sleep analysis, demonstrating the potential for AI to exceed human performance in clinical settings.

Key words: sleep staging; interrater reliability; deep learning; transfer learning; confidence; scorability; home sleep recordings

Submitted for publication: February 20, 2024; **Revised:** August 4, 2024

Graphical Abstract



Statement of Significance

This work represents an important milestone in deployable sleep analysis, with our automated system surpassing human experts through targeted deep learning. We demonstrate state-of-the-art performance by pretraining a model on diverse polysomnography data and then fine-tuning it to home sleep studies. Notably, model-expert agreement exceeds inter-expert agreement across various scorability scores, which reflect both signal quality and model confidence. To the best of our knowledge, this is the first model to conclusively outperform human review for sleep staging. By covering real-world training strategies, direct clinician benchmarking, and reliability assessments, our study defines best practices for clinical validation of AI in sleep medicine. The presented system paves the way for broad access to accurate sleep analysis, with implications for precision diagnostics and treatments.

Sleep is a fundamental physiological process that plays a crucial role in maintaining health [1, 2] Sleep serves numerous restorative processes, including memory consolidation, immune system regulation, and hormonal balance [3]. Consequently, accurate analysis of sleep has implications for diagnosing and treating various medical conditions, including sleep, neurological, and mental health disorders [4, 5] A critical facet of sleep analysis is sleep staging: categorization of different sleep stages based on characteristic electroencephalographic (EEG) patterns, eye movements, muscle tone, and other physiological signals [6–8].

The current gold standard for sleep staging is the manual analysis of polysomnography (PSG) recordings following American Academy of Sleep Medicine (AASM) guidelines [6]. These guidelines classify each 30-second epoch into wake (W), rapid eye movement (REM), and non-REM stages 1 to 3 (N1, N2, and N3). However, manual sleep staging methods are labor-intensive and time-consuming. Inter-expert variability is problematic [9, 10], with agreement ranging from 60% between experts from different institutions to 80% within the same institution. Studies have investigated interrater reliability (IRR) in sleep staging, revealing that variability is primarily due to epochs that are difficult to classify and that agreement varies by scorer, diagnosis, and record, with inter-laboratory variability often exceeding that within a single laboratory [11–14]. AI-based automated sleep staging is a potential solution.

Impressive automated sleep staging has been achieved in multiple recent publications by utilizing large expert-annotated datasets to train neural network models, including extreme learning machines [15], convolutional neural networks (CNNs) [16, 17], CNNs + recurrent neural networks [18–24], and graph neural networks [25–28] However, performance can suffer when applying a sleep staging neural network trained on one population to a new population.

Failure to generalize can arise from biological differences (e.g. different distribution of pathology, race, sex, or medication use), or technical differences (e.g. different signal quality, montage, or recording equipment). Training a new neural network for each new cohort is infeasible because scoring 1000s of new PSGs is labor-prohibitive. An effective solution in other domains has been *TL*: fine-tuning (calibrating) a model trained on a large set of data using a small number of labeled samples from the new domain. This is done by holding constant or freezing most neural network parameters while adjusting a small number of parameters, typically from the final few layers of the network [29–32].

Transfer learning (TL) has been applied in sleep staging before—for instance [33] introduced RobustSleepNet to address

PSG montage variability [34] utilized deep TL to mitigate data variability and inefficiency [35] adapted ECG-trained models to photoplethysmography for home monitoring, and Van et al [36] examined the performance differences between pretraining, training on a new target dataset only, and applying transfer learning.

In this study, we build on these works by (1) investigating different TL approaches to determine which strategies yield better results, (2) utilizing a dataset independently scored by three experts, enabling us to directly measure TL success through a comparison of ML-expert to expert-expert agreement, a novel aspect of our research, and (3) extending sleep staging research with a new scorability model to assess output reliability. This comprehensive approach ensures that our findings are robust and applicable across various settings and model architectures, thereby advancing practical sleep staging in clinical care.

Specifically, we applied a base model trained on a diverse set of PSGs from the United States of America to a new international PSG dataset from the Korean Genome and Epidemiology Study (KoGES). KoGES PSGs present technical challenges (only one EEG electrode compared to at least two for base model training) and biological challenges (demographic differences since the base model relied heavily on Western/Caucasian cohorts, while the KoGES cohort is Korean). As expected, the base model performs less well on KoGES data than on the model development cohort. We hypothesized that TL with a small number of labeled PSGs could improve the base model to exceed human reliability. To test this, we recruited multiple experts to score KoGES PSGs. We examined various strategies for TL to calibrate the base model, employing new approaches for strategically using the training data. We demonstrate that TL improves the base model such that it exceeds the reliability of experts.

We noticed that expert staging reliability varied across PSGs, and wondered if this reflected difficulties from poor signal data quality in these unattended polysomnograms. It was therefore possible that TL only outperforms experts on certain cases, which we hypothesized would be cases with high signal quality. To test this, we developed an automated "scorability index." We find that expert reliability is indeed strongly dependent on signal quality. Contrary to expectation, the calibrated model outperformed experts even on signals with low scorability.

Methods

Sleep staging base model

As a base model, we used a previously published neural network, ProductGraphSleepNet [25] with the following layers: spatial attention (SpAtt), product graph learning (PGL), attentive graph convolutional (AGC), bidirectional gated recurrent unit (BiGRU), graph-wise attention network (GwAT), and a fully connected layer. Channels included as inputs to the base model ("base model input signals") include two EEG (C3-M2 and C4-M1 or Cz-Oz), one electrooculogram (EOG, E1-M2), chin electromyogram (EMG), abdominal and thoracic respiratory effort, and electrocardiogram (ECG).

To select the base model for our sleep staging task, we conducted a thorough literature review to identify state-of-the-art models known for their strong performance in sleep analysis. We evaluated five leading models, including a hybrid CNN and LSTM model by Biswal et al. [28], SleepFCN by Goshtasbi et al. [37], U-Sleep by Perslev et al. [17], U-NET by Zhang et al. [38], and ProductGraphSleepNet by Einizade et al. [25]. These models were trained from scratch on consistent datasets, and their performance was compared using PSG recordings from 8000 individuals divided into training, validation, and test sets. ProductGraphSleepNet emerged as the top performer.

To ensure the TL studies were broadly applicable, we also employed SleepNet, a model combining CNN and LSTM architectures introduced by Biswal et al. [28] This model uses CNNs to extract spatial features from raw EEG signals and RNNs to capture temporal dependencies across sleep epochs. We tested this model with six EEG signals from the MGH cohort, using spectrogram representations and applying the Adam optimizer with a learning rate of 10^{-4}. The results from this model were compared to ensure consistency with the primary base model, ProductGraphSleepNet.

Datasets used to train the base model

The base model was trained with a large (N = 21 764 participants) multicenter dataset that included PSGs from one clinical sleep center (MGH) and three large epidemiological cohorts (SHHS, MESA, and MrOs), Table 1. These data were recorded between 10 (respiratory signals) and 512 Hz (ECG and EEG signals) and resampled to 200 Hz for analysis.

Dataset used for TL

KoGES.

The KoGES dataset included PSGs from 800 participants with EEG channel C4-M1, one EOG E2-M1, chin EMG, abdominal and thoracic respiratory effort, and ECG. To match the number of channels required by the base model, we duplicated the EEG channel. We resampled all signals to 200 Hz.

Six hundred and forty-eight out of the KoGES recordings were staged by a single expert, and 152 were scored independently by three experts. All experts are AASM-certified sleep technologists working in the Beth Israel Deaconess Medical Center sleep laboratory.

The triple-scored KoGES data was used to evaluate model performance.

Base model training

Preprocessing consisted of notch filtering (60 Hz) and bandpass filtering with the following passbands: EEG, EOG, and chin EMG: [0.1–25 Hz]; thoracic and abdominal effort: [0–10 Hz]; and ECG [0.3–40 Hz]. The filter settings were chosen based on a priori knowledge and preliminary analyses, which suggested that these specific ranges best capture the salient features for effective sleep stage classification by our model. The model follows the AASM sleep stage standard, which includes stages W, N1, N2, N3, and REM. When data was originally scored using the Kales and Rechtschaffen manual, stages S3 and S4 were merged into a single stage labeled as N3.

We trained the model with data from all cohorts (MGH, SHHS, MROS, and MESA), with 70% allocated for training, 15% for validation, and 15% for testing within each cohort, ensuring no overlap of participants. This process leverages PSG data from 25 749 participants to develop the final sleep staging model that we used as our base model.

During model training, we used an early stopping mechanism, which monitored the validation Sørensen–Dice coefficient [17, 39]. We configured a patience value of 100 training epochs. Optimization was done using the Adam optimizer with a learning rate of 10^{-5} [40].

Table 1. Demographics

	MGH	SHHS	MESA	MrOS	KoGES single-scored	KoGES triple-scored
N participants	14 859	5793	2055	2898	648	152
Age (mean, IQR)	52 (41, 65)	63 (55, 72)	68 (62, 76)	76 (72, 80)	61 (56, 65)	62 (57, 64)
N (%) age 0–20	507 (3)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
N (%) age 20–40	2991 (20)	6 (0)	0 (0)	0 (0)	0 (0)	0 (0)
N (%) age 40–60	6147 (41)	2291 (40)	331 (16)	0 (0)	348 (54)	70 (46)
N (%) age 60–80	4775 (32)	3098 (54)	1399 (68)	2085 (72)	294 (45)	80 (53)
N (%) age 80–100	438 (3)	311 (5)	274 (13)	683 (24)	6 (1)	2 (1)
N (%) sex female	6396 (43)	3033 (52)	1102 (54)	0 (0)	315 (45)	67 (44)
Race/Ethnicity*						
N (%) Asian	486 (3)	N/A	250 (12)	85 (3)	648 (100)	152 (100)
N (%) black	913 (6)	514 (9)	571 (28)	99 (3)	0	0
N (%) white	11390 (75)	4899 (85)	743 (36)	2638 (91)	0	0
N (%) Hispanic	N/A	N/A	491 (24)	N/A	0	0
N (%) other	1845 (12)	380 (7)	0 (0)	76 (3)	0	0
Type cohort	Sleep laboratory attended PSG	Community-dwelling, home-recording	Community-dwelling, home-recording	Community-dwelling, home-recording	Community-dwelling, home-recording	Community-dwelling, home-recording

*MESA collected race "Chinese," not "Asian." MESA collected "Hispanic" as part of the race questionnaire. All KoGES participants lived in South Korea and identified as "Northeast Asian."

It is important to note that the evaluation of the base model on the development cohort is outside the present study. Here we focus on evaluating and improving the base model using TL in the KoGES dataset.

TL experiments

Here we describe our strategy for transferring knowledge from the large multi-cohort training cohort to KoGES PSGs using a small number of training samples. The transfer is achieved by freezing specific layers within the pretrained base model and fine-tuning the remaining layers using annotated PSGs from KoGES. In our base model, ProductGraphSleepNet [25], we froze the Spatial Attention layer, PGL, AGC, and BiGRU layers, and trained the remaining layers until reaching the stopping criterion. Subsequent training exclusively fine-tuned the last two layers: the GwAT layer and fully connected layers. The same early-stopping strategy used for base model training was used for TL training; however, due to the limited number of participants, fine-tuning was stopped when the validation Sørensen–Dice coefficient showed no improvement for 30 consecutive fine-tuning epochs.

To investigate how different TL choices influence model performance on the evaluation dataset (triple-scored KoGES dataset), we conducted 6 experiments:

- Experiment 1: no fine-tuning (use base model).
- Experiment 2: TL with a random subset of 324 single-scored KoGES PSGs.
- Experiment 3: TL with 324 single-scored KoGES PSGs that were "difficult." These samples had lower scorability scores (see below) than the second half of the single-scored dataset.
- Experiment 4: TL with all 648 single-scored KoGES PSGs.
- Experiment 5: TL with half of the triple-scored KoGES PSGs in a 2-fold cross-validation approach (76 for training + internal validation, 76 for testing). To train the model, the majority vote of the three scorers for each epoch was used (with a random selection of the three votes if the three votes were all different).
- Experiment 6: TL with two steps, first using all 648 single-scored KoGES PSGs (as in experiment 4) and subsequently 2-fold cross-validation learning with the triple-scored PSGs (as in experiment 5).

Model evaluation

To assess model performance, we used the following metrics:

1. Confusion matrix: offers a detailed breakdown of how well two raters agree on any given sleep stage. For triple-scored PSGs, we compute confusion matrices to assess expert-expert and expert-model agreement.
2. Cohen's Kappa: a summary statistic for agreement between two raters. We compute Cohen's Kappa both among pairs of experts and between experts and the model.
3. Receiver operating characteristics (ROC) curve and precision-recall curve (PRC): The ROC and PRC analyses are conducted in a stage-wise manner by consolidating expert labels and model probabilities into binary categories per sleep stage. This enables fine-grained assessment of model capabilities by stage. To compute these metrics, we treat each expert as ground truth and compare model performance to the agreement level between experts. EUC and PRC analyses provide insights into whether the model reaches human-level performance.

We introduce a metric named "Experts Under the Curve (EUC)," which quantifies the percentage of experts' Receiver Operating Characteristic (ROC) or Precision-Recall Curve (PRC) operating points that are surpassed by the model's curve. Specifically, for

each expert rater chosen as the ground truth, other raters are evaluated against this reference, generating distinct operating points on the curve. Simultaneously, the model generates its own ROC curve based on predictions versus the same ground truth. This process is repeated for each expert as ground truth, thus collecting multiple sets of operating points and ROC curves. The EUC metric is then calculated by determining the proportion of these expert operating points that fall below the model's respective ROC curves. This method effectively measures how often the model's performance matches or exceeds the inter-expert agreement levels across different comparisons.

Scorability model

The scorability model integrates information from the sleep data and the model output (but not the expert annotations) to generate a dimensionless score from –1 to 1 representing the expected agreement between the model and experts.

The scorability model used sleep features, signal features, as well as model output confidence features. Sleep features include metrics such as sleep efficiency, sleep fragmentation indices, and percentages of various sleep stages calculated from the model-predicted hypnogram. Signal features are computed from the EEG, EOG, Chin EMG, and respiratory effort signals, capturing key parameters such as signal-to-noise ratio and abiological amplitude fluctuations. Additionally, power spectral densities are calculated within specific EEG frequency bands—delta, theta, alpha, and beta—to analyze brain activity relevant to different sleep stages. To represent the model's confidence in its sleep stage classifications, we compute the entropy of the predicted sleep stage distributions.

The model employed an Elastic Net linear regression [41] approach, combining the penalties of both ridge and lasso regression to enhance model robustness and prevent overfitting. For training, the model used data from the single-scored KoGES PSGs, aiming to predict the Cohen's Kappa statistic, a measure of interrater reliability between the model's predictions and expert annotations. The training and evaluation process employed a 10-fold nested cross-validation scheme, ensuring that predictions are made on independent, left-out samples. In each fold, optimal Elastic Net parameters, such as the regularization strength and the l1-ratio, are independently tuned on the validation subset of the training data to maximize model performance before evaluation on unseen test data [42, 43]. After the cross-validation process, a final model was trained using the same approach but incorporating all available training data. This final model was then applied to the triple-scored data, and the resulting model equation was reported.

Results

TL experiments

Figure 1 shows the interrater analysis results on the triple-scored KoGES PSGs ($N = 152$). The mean expert agreements (Cohen's Kappa) were 0.60 (expert 1 vs expert 2), 0.64 (expert 1 vs expert 3), and 0.61 (expert 2 vs expert 3). The base model showed significantly lower mean agreement with experts compared to the expert–expert agreement (Cohen's Kappa 0.52 vs 0.64). Calibration using 324 randomly selected single-scored KoGES PSGs (experiment 2) led to a model with similar performance to the experts. Fine-tuning experiments 3–6 resulted in models

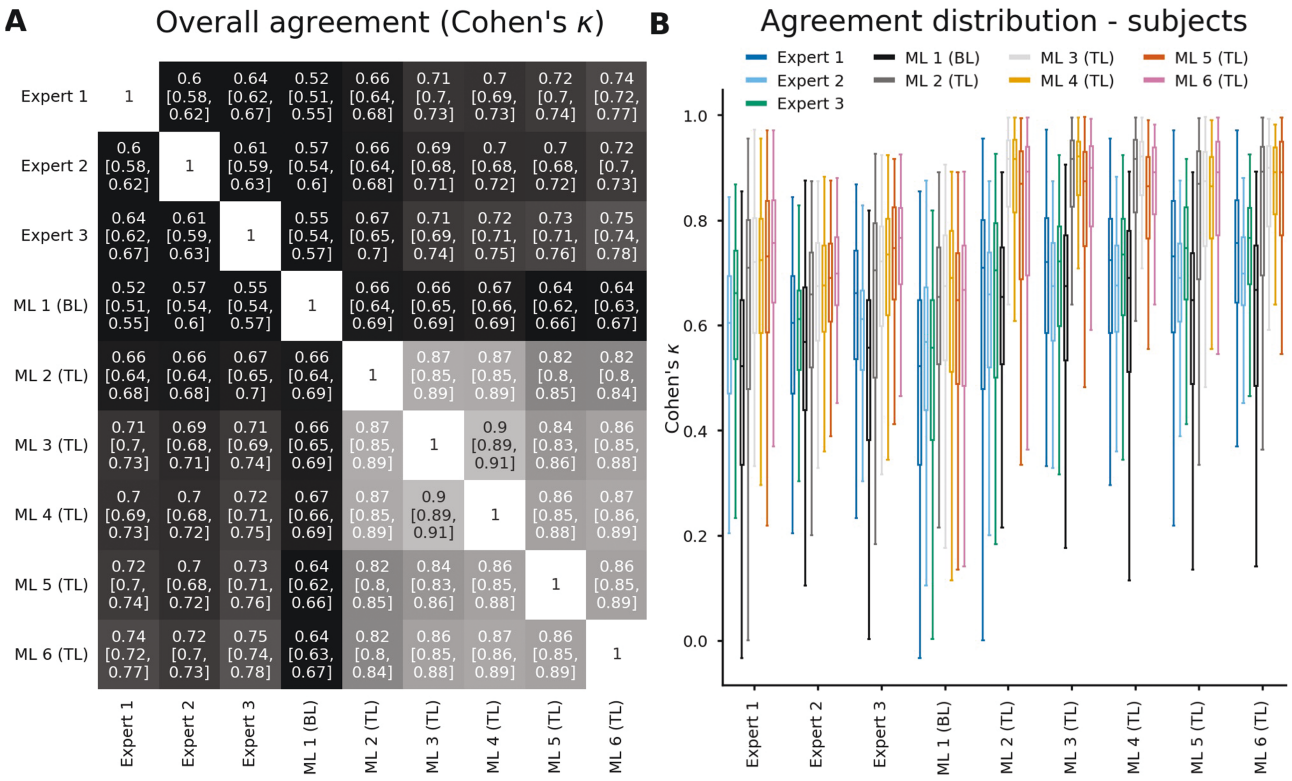


Figure 1. Interrater analysis results on triple-scored KoGES PSGs ($N = 152$). (**A**) The agreement matrix depicts the agreement between pairs of two raters, quantified by mean Cohen's Kappa (and 95% confidence interval). The raters include the three human experts, the base machine learning model before fine-tuning, and the five machine learning models from the fine-tuning experiments. (**B**) The distribution of Cohen's Kappa across all participants. All raters (experts and models) are considered as "ground truth" in turn.

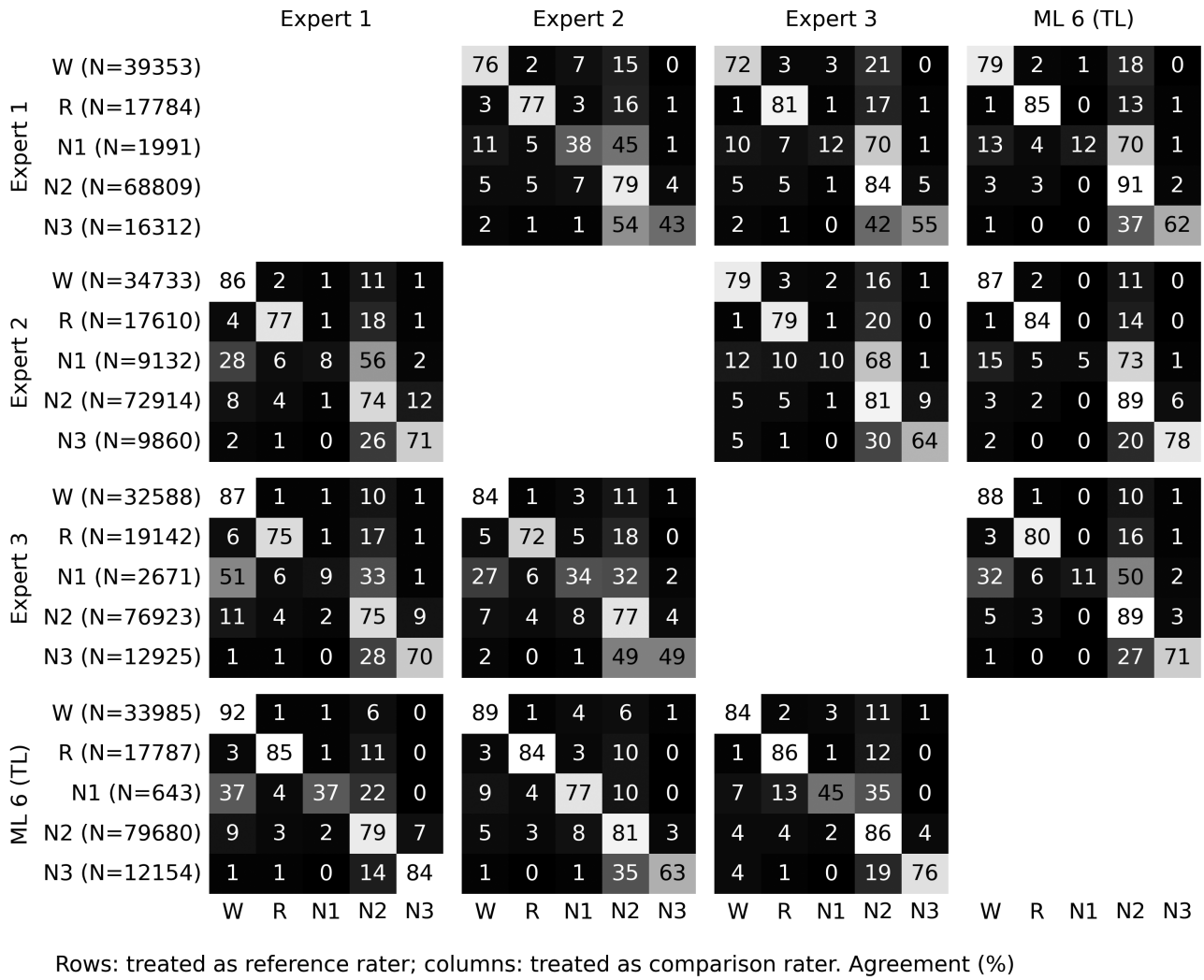


Figure 2. Pairwise confusion matrices between experts and the best-performing transfer learning model (experiment 6). Raters in the rows are considered "ground truth" while raters in the columns represent "comparative evaluators." Percentages indicate agreement with the ground truth per sleep stage. Numbers in parentheses show epochs per stage.

with progressively higher model-expert agreement compared with expert–expert agreement. The 2-step fine-tuning approach (experiment 6) yielded the highest reliability values with model-expert agreements of 0.74 (expert 1), 0.72 (expert 2), and 0.76 (expert 3).

In Figure 2 and Figure 3, we investigate the experiment-6 model more closely. Figure 2 shows pairwise classwise agreement matrices. Figure 3A shows the distribution of Cohen's Kappa as boxplots. Figure 3B shows ROC and PRC curve analyses. The mean ROC AUCs are between 0.84 (stage N1) and 0.96 (stages W and R), while mean PRC AUCs are between 0.24 (stage N1) and 0.91 (stage W). Stagewise ROC and PRC experts-under-the-curve analyses yielded values of 1, demonstrating that the model not only showed increased overall reliability compared to expert–expert agreement, but this was also true for every sleep stage both in the sensitivity-specificity trade-off (ROC) and in precision-sensitivity trade-off (PRC).

We observed similar performance when evaluating the alternative deep neural network architecture (CNN + LSTM), with results slightly lower than those of the graph neural network. The 95% confidence intervals for various performance metrics often overlapped. This indicates that the results are not due to a particular model architecture choice alone but reflect general deep learning behavior in sleep staging.

Scorability model

The final scorability model included 60 non-zero coefficients, with the most influential variables being the predicted sleep fragmentation index, entropy values for wake, N2, N3, and REM stage predictions, predicted REM latency, and the standard deviation of the EEG signal during the predicted wake. All signal modalities, including EEG and non-EEG signals such as EOG, Chin EMG, and respiratory effort, contributed to the model. All coefficients of the model are listed in Supplementary Material.

The cross-validation results showed that the explained variance (R-squared value) of the observed agreement between the model and experts (Cohen's Kappa) by the scorability scores was $R^2 = 0.42$ (Figure 4A). The root mean squared error (RMSE) between observed and predicted Kappa values was 0.14. A sensitivity analysis, excluding any outliers outside the range of one and a half times the interquartile range below the first quartile and above the third quartile (removing 32 outliers), resulted in $R^2 = 0.34$ and RMSE = 0.14.

After applying this model to the triple-scored data, we plotted agreement between raters versus the scorability scores per (3

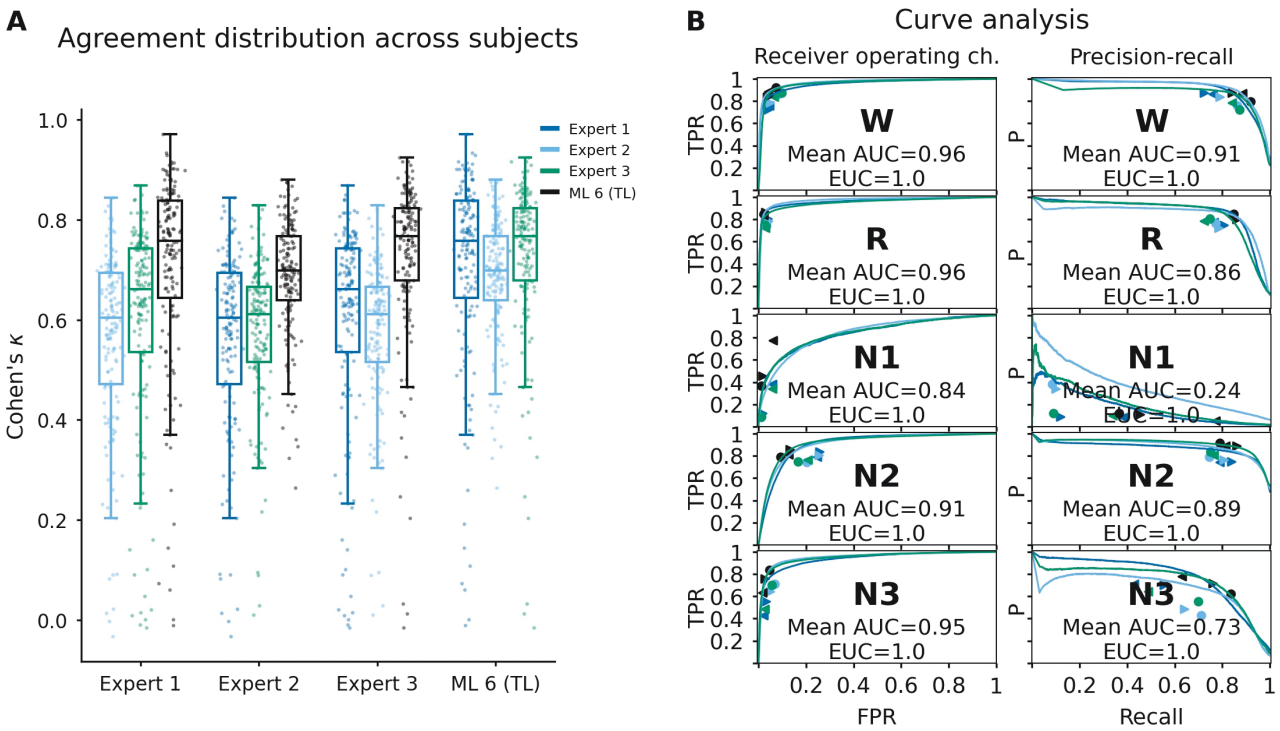


Figure 3. (**A**) Distribution of Cohen's Kappa, depicted as boxplots with individual data points. (**B**) ROC and PRC curve analyses. Mean ROC AUCs range from 0.84 (stage N1) to 0.96 (stages W and R), while mean PRC AUCs range from 0.24 (stage N1) to 0.91 (stage W). Stagewise ROC and PRC experts-under-the-curve analyses consistently yield values of 1, indicating the model's significantly increased overall agreement compared to expert–expert agreement. This holds true for every sleep stage in both sensitivity-specificity tradeoff (ROC) and precision-sensitivity tradeoff (PRC).

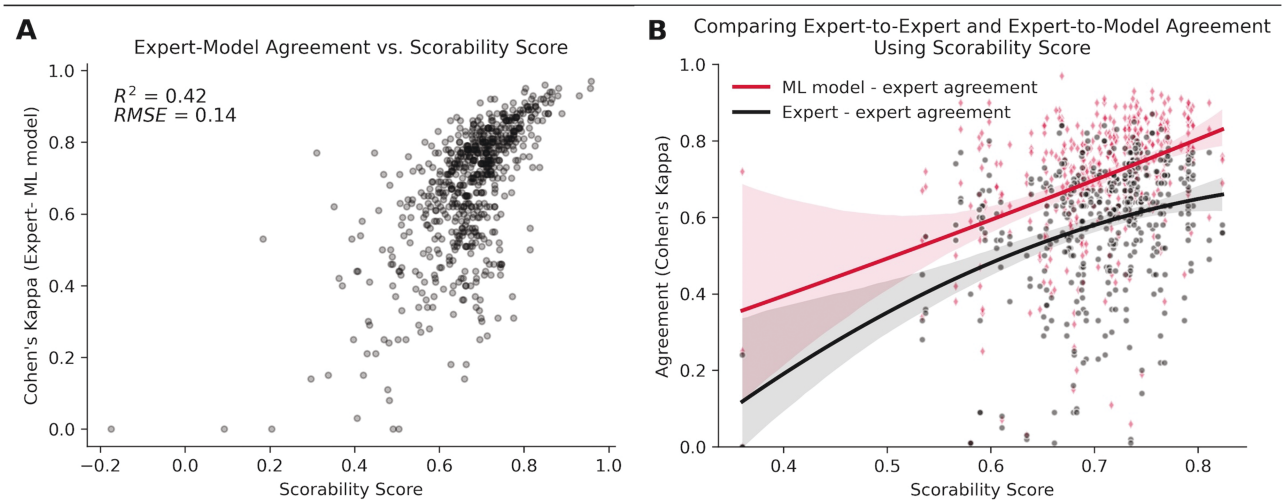


Figure 4. (**A**) The scorability model achieved an explained variance (R^2) of 0.42 and a root mean squared error of 0.14 in predicting the agreement (Cohen's Kappa) between the model and experts (N = 648 participants). (**B**) Agreement between experts and machine learning model stratified by scorability index for the triple-scored data (N = 152). Scatter plots show agreement (Cohen's kappa) between pairs of experts (circles) and between the machine learning model and individual experts (diamonds) on the y-axis versus the scorability index on the x-axis for each PSG. Solid lines indicate quadratic fits for inter-expert agreement and model-expert agreement with 95% confidence intervals shaded. Recordings with scorability ≥ 0.55 demonstrated significantly higher model-expert agreement than inter-expert agreement. The scorability index, combining automated EEG signal quality and model confidence assessments, accurately predicts the reliability of the model sleep staging. This enables quality control thresholds to determine which recordings have suitably high accuracy for clinical or research use. The figure highlights how targeted transfer learning allows deep neural networks to match or surpass human experts for automated sleep analysis on all signal- and model confidence levels.

data points for the three expert-expert agreements, 3 data points for model-expert agreements). We fit second-order linear regression models to both the expert-expert and model-expert data and observed significantly higher model-expert agreement compared to expert–expert agreement for scorability scores of 0.55 or higher, see Figure 4B, whereas confidence intervals overlapped for scorability scores below 0.55. Together, the scorability results show that (1) the expected agreement between the model and experts can be inferred with moderately good accuracy from scorability scores and (2) for PSGs with scorability 0.55 or higher, the scoring

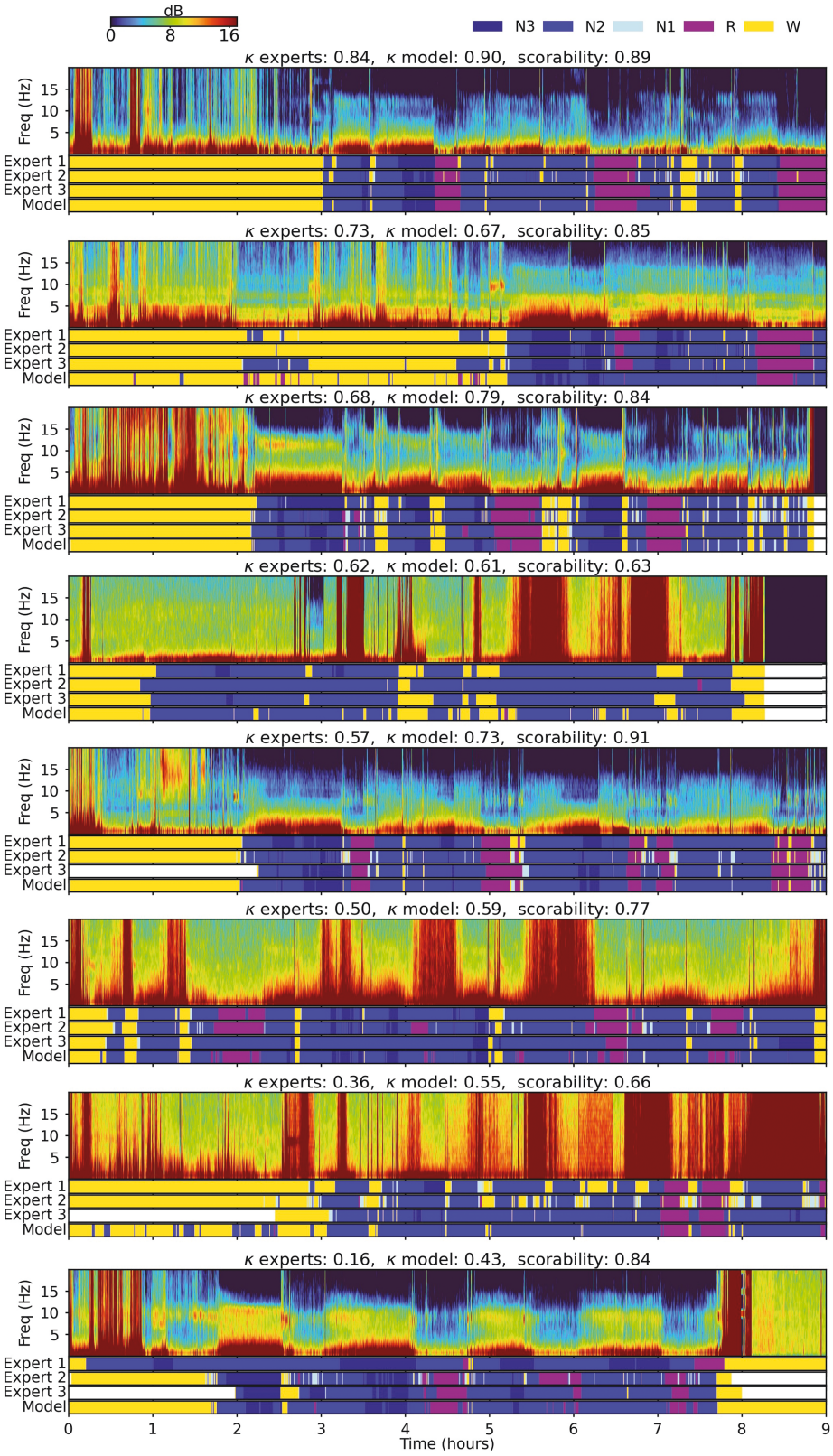


Figure 5. Eight sample recordings from the triple-scored KoGES PSGs, including sleep stage annotations produced by the three experts and the experiment 6 machine learning model. The mean agreement between experts, model-expert agreement, and scorability scores are indicated at the top of each EEG spectrogram. The selected samples span the distribution of kappa values among experts, showcasing a spectrum of recordings with varying levels of scorabilty and agreement.

produced by the machine learning model is expected to have a higher agreement with experts than experts have among themselves, whereas for recordings with scorability less than 0.55, there is no expected difference between model and expert agreement. Further, scorability scores allow us to predict agreement between the model and experts for data not scored by any expert. For example, a scorability value of 0.55 predicts a mean kappa value of 0.71, while a scorability score of 0.80 predicts a mean kappa value of 0.76. As one increases the desired threshold (and therefore expected agreement), the proportion of recordings that lie above the threshold decreases (e.g. while 98% of the data lie above scorability 0.55, only 22% lie above 0.75). Scorability can therefore inform about the expected reliability of an automated sleep staging output and can help determine which data should be trusted in clinical practice or research analyses.

Figure 5 shows eight sample recordings as an EEG spectrogram and the accompanying expert and model sleep stage annotations.

Discussion

This study demonstrates the potential of strategically integrating pretrained deep neural networks with targeted TL to exceed human expert reliability in automated sleep staging on new, unseen data with different demographics and technical details compared to the training data of the base model. Our calibrated model not only matched but surpassed human interrater reliability (based on assessments by three experts) across all sleep stages, as evidenced by Cohen's Kappa and the areas under the ROC and PRC curves within our dataset. This result is particularly notable given the model's initial underperformance with a new cohort that differed demographically and technically from the training data.

The practical implications of these findings are significant. Prior to this study, there was a wide consensus supported by substantial data that automated sleep staging could transform clinical settings by enhancing the efficiency, reproducibility, and accessibility of sleep diagnostics [44]. This technology is particularly promising for remote sleep studies, where it could facilitate more accurate and efficient data analysis without the need for human expertise, which is in short supply.

Despite our initial model being trained on multiple large-scale cohorts, applying it "out of the box" to a new dataset with different demographics and recording technologies resulted in a lower agreement between the ML model and experts compared to expert–expert agreement. With the application of different TL techniques, two key outcomes were observed: the mean ML-expert agreement surpassed the expert–expert agreement, indicating the model performed better than experts, and TL was more effective with a strategic choice of data type and increased data volume.

Therefore, our study adds nuance to the scientific and clinical aim of using automated tools established via machine learning in clinical practice. To ensure that a given ML model performs as well as trained experts, a similar approach to ours, involving both model fine-tuning (TL) and model evaluation (comparing model-expert agreement to expert-expert agreement), can be used. We ensured that these TL results were not only applicable to the primary architecture chosen (graph neural network) but also held true for more traditional deep learning architectures (CNN and LSTM).

The scorability model provides a method to predict the expected agreement between automated sleep staging systems and expert annotations using only the model output and sleep recording data, without needing prior expert scoring. Higher scorability indices, which suggest better signal quality and greater confidence in the model's output, correlated with better agreement between the model and experts. This functionality is beneficial in clinical settings, as it allows for the assessment of the reliability and variability of model outputs before involving expert review. By enabling the identification of high-confidence recordings, the scorability model can help optimize the use of expert time and streamline clinical workflows.

Our study has some limitations. The space of deep learning architectures and TL strategies is vast. While we believe our experimental choices are reasonable, there are more parameters to explore. Additionally, the current study's cohort was relatively homogeneous, collected in a single country using the same equipment. The TL results and scorability analysis need to be replicated in more diverse datasets.

Future research should focus on expanding this technology's applicability to more diverse datasets and clinical settings. Investigating the reasons behind human scoring variability could also enhance the reliability of automated scoring systems. Additionally, examining the ethical and quality control considerations of entirely automated sleep staging in clinical settings is essential as this technology advances.

In conclusion, the integration of pretrained deep neural networks and TL in our study produced an automated sleep staging system that surpasses human expert reliability. Moreover, we have demonstrated that model-expert agreement can be modeled at least moderately well based on features from biosignals and model outputs. This advance holds promise for enhancing the research reproducibility, global accessibility of sleep diagnostic tools, and usability in clinical practice, marking a significant step forward in the field of sleep medicine.

Supplementary material

Supplementary material is available at *SLEEP* online.

Disclosure Statement

Financial disclosure: This work was supported by grants from the NIH (R01NS102190, R01NS102574, R01NS107291, RF1AG064312, RF1NS120947, R01AG073410, R01HL161253, R01NS126282, and R01AG073598) and NSF (2014431). This study was also supported by the Korea Disease Control and Prevention Agency (grant nos. 2011-E71004-00, 2012-E71005-00, 2013-E71005-00, and 2014-E71003-00); Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (NRF-2020R1I1A1A01071011 and NRF-2022R1I1A1A01065700); the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health & Welfare (no. HI20C0469). *Nonfinancial disclosure:* Dr. Westover is a co-founder, serves as a scientific advisor and consultant to, and has a personal equity interest in Beacon Biosignals. R.J. Thomas discloses (1) patent and license/royalties from MyCardio, LLC, for the ECG-spectrogram; (2) patent and license/royalties from DeVilbiss-Drive for an auto-CPAP algorithm; and (3) consulting for Guidepoint Global and GLG Councils.

Data Availability

Individual cohort data used in this study are stored in the National Biobank of Korea Database and managed by the National Institute of Health, Republic of Korea (https://biobank.nih.go.kr). Further information and requests for polysomnographies would

be via e-mail to chol-shin@korea.ac.kr. In response to reasonable requests for this data, we could share the summary of data and results after blinding personal information under the approval of the Distribution Review Board in NIH, Republic of Korea.

The code used for model training and evaluation, along with the resulting data, is available at https://github.com/bdsp-core/sleep_staging_tl_scorability

References

1. Worley SL. The extraordinary importance of sleep: the detrimental effects of inadequate sleep on health and public safety drive an explosion of sleep research. *P T*. 2018;**43**(12):758–763. doi: 10.5664/jcsm.7176
2. Stickgold R. Sleep-dependent memory consolidation. *Nature*. 2005;**437**(7063):1272–1278. doi: 10.1038/nature04286
3. Vyazovskiy V. Sleep, recovery, and metaregulation: explaining the benefits of sleep. *Nat Sci Sleep*. 2015;**171**:171. doi: 10.2147/nss.s54036
4. Mahowald MW, Schenck CH. NREM SLEEP PARASOMNIAS. *Neurol Clin*. 1996;**14**(4):675–696. doi: 10.1016/s0733-8619(05)70280-2
5. Malow BA, Foldvary-Schaefer N, Vaughn BV, *et al*. Treating obstructive sleep apnea in adults with epilepsy: a randomized pilot trial. *Neurology*. 2008;**71**(8):572–577. doi: 10.1212/01.wnl.0000323927.13250.54
6. Berry RB, Budhiraja R, Gottlieb DJ, *et al*. Rules for scoring respiratory events in sleep: update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events: Deliberations of the Sleep Apnea Definitions Task Force of the American Academy of Sleep Medicine. *J Clin Sleep Med*. 2012;**08**(05):597–619. doi: 10.5664/jcsm.2172
7. Silber MH, Ancoli-Israel S, Bonnet MH, *et al*. The visual scoring of sleep in adults. *J Clin Sleep Med*. 2007;**03**(02):121–131. doi: 10.5664/jcsm.26814
8. Rosenberg RS, Van Hout S. The American Academy of Sleep Medicine Inter-scorer Reliability Program: sleep stage scoring. *J Clin Sleep Med*. 2013;**09**(01):81–87. doi: 10.5664/jcsm.2350
9. Nasiri S, Clifford GD. Attentive adversarial network for large-scale sleep staging. In: *Proceedings of the 5th Machine Learning for Healthcare Conference, in Proceedings of Machine Learning Research, PMLR* 2020;**126**:457–478.
10. Nasiri S, Clifford GD. Boosting automated sleep staging performance in big datasets using population subgrouping. *Sleep*. 2021;**44**(7). doi: 10.1093/sleep/zsab027
11. Basner M, Griefahn B, Penzel T. Inter-rater agreement in sleep stage classification between centers with different backgrounds. *Somnologie*. 2008;**12**(1):75–84. doi: 10.1007/s11818-008-0327-y
12. Danker-Hopfe H, Kunz D, Gruber G, *et al*. Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders. *J Sleep Res*. 2004;**13**(1):63–69. doi: 10.1046/j.1365-2869.2003.00375.x
13. Norman RG, Pal I, Stewart C, Walsleben JA, Rapoport DM. Interobserver agreement among sleep scorers from different centers in a large dataset. *Sleep*. 2000;**23**(7):901–908.
14. Younes M, Raneri J, Hanly P. Staging sleep in polysomnograms: analysis of inter-scorer variability. *J Clin Sleep Med*. 2016;**12**(06):885–894. doi: 10.5664/jcsm.5894
15. Sun H, Jia J, Goparaju B, *et al*. Large-scale automated sleep staging. *Sleep* 2017;**40**(10):zsx139. doi: 10.1093/sleep/zsx139
16. 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 Trans Neural Syst Rehabil Eng*. 2018;**26**(4):758–769. doi: 10.1109/TNSRE.2018.2813138
17. Perslev M, Darkner S, Kempfner L, Nikolic M, Jennum PJ, Igel C. U-Sleep: resilient high-frequency sleep staging. *NPJ Digit Med*. 2021;**4**(1):72. doi: 10.1038/s41746-021-00440-5
18. 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*. 2017;**25**(11):1998–2008. doi: 10.1109/TNSRE.2017.2721116
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. *Biomed Signal Proc Control*. 2020;**61**:102037. doi: 10.1016/j.bspc.2020.102037
20. Dong H, Supratak A, Pan W, Wu C, Matthews PM, Guo Y. Mixed neural network approach for temporal sleep stage classification. *IEEE Trans Neural Syst Rehabil Eng*. 2018;**26**(2):324–333. doi: 10.1109/TNSRE.2017.2733220
21. Biswal S, Sun H, Goparaju B, Westover MB, Sun J, Bianchi MT. Expert-level sleep scoring with deep neural networks. *J Am Med Informat Assoc*. 2018;**25**(12):1643–1650. doi: 10.1093/jamia/ocy131
22. Vallat R, Walker MP. An open-source, high-performance tool for automated sleep staging. *Elife*. 2021;**10**:e70092. doi: 10.7554/eLife.70092
23. Phan H, Chén OY, Tran MC, Koch P, Mertins A, De Vos M. "XSleepNet: multi-view sequential model for automatic sleep staging.". *IEEE Trans Pattern Anal Mach Intell*. 2021;**44**(9):1–1. doi: 10.1109/tpami.2021.3070057
24. Olesen AN, Jørgen Jennum P, Mignot E, Sorensen HBD. Automatic sleep stage classification with deep residual networks in a mixed-cohort setting. *Sleep*. 2021;**44**(1). doi: 10.1093/sleep/zsaa161
25. Einizade A, Nasiri S, Sardouie SH, Clifford GD. ProductGraphSleepNet: sleep staging using product spatio-temporal graph learning with attentive temporal aggregation. *Neural Netw*. 2023;**164**:667–680. doi: 10.1016/j.neunet.2023.05.016
26. Song T, Zheng W, Song P, Cui Z. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Trans Affective Comput*. 2020;**11**(3):532–541. doi: 10.1109/taffc.2018.2817622
27. Zhang T, Wang X, Xu X, Chen CLP. GCB-Net: graph convolutional broad network and its application in emotion recognition. *IEEE Trans Affective Comput*. 2022;**13**(1):379–388. doi: 10.1109/taffc.2019.2937768
28. Congedo M, Barachant A, Bhatia R. Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review. *Brain Comp Inter*. 2017;**4**(3):155–174. doi: 10.1080/2326263x.2017.1297192
29. Wang P, Li J, Wang S, Zhang F, Shi J, Shen C. A new meta-transfer learning method with freezing operation for few-shot bearing fault diagnosis. *Meas Sci Technol*. 2023;**34**(7):074005.
30. Alayrac J-B, Donahue J, Luc P, *et al*. "Flamingo: a visual language model for few-shot learning". *Adv Neural Inform Process Syst*. 2022;**35**:23716–23736.
31. Wang Y, Yao Q, Kwok JT, Ni LM. Generalizing from a few examples: a survey on few-shot learning. *ACM Comput Surv*. 2021;**53**(3):1–34. doi: 10.1145/3386252
32. Kim H, Seo P, Byun J-I, Jung K-Y, Kim KH. Spatiotemporal characteristics of cortical activities of REM sleep behavior disorder revealed by explainable machine learning using 3D convolutional neural network. *Sci Rep*. 2023;**13**(1):8221. doi: 10.1038/s41598-023-35209-1
33. Guillot A, Thorey V. RobustSleepNet: transfer learning for automated sleep staging at scale. *IEEE Trans Neural Syst Rehabil Eng*. 2021;**29**:1441–1451. doi: 10.1109/TNSRE.2021.3098968
34. Phan H, Lorenzen KP, Heremans E, *et al*. "L-SeqSleepNet: Whole-cycle long sequence modelling for automatic sleep staging". *IEEE J Biomed Health Inf*. 2023;**27**:4748–4757. doi: 10.1109/jbhi.2023.3303197

35. Radha M, Fonseca P, Moreau A, *et al*. A deep transfer learning approach for wearable sleep stage classification with photoplethysmography. *NPJ Digit Med*. 2021;**4**(1):135. doi: 10.1038/s41746-021-00510-8
36. Van Der Aar JF, Van Den Ende DA, Fonseca P, *et al*. Deep transfer learning for automated single-lead EEG sleep staging with channel and population mismatches. *Front Physiol*. 2024;**14**:1287342. doi: 10.3389/fphys.2023.1287342
37. Goshtasbi N, Boostani R, Sanei S. SleepFCN: a fully convolutional deep learning framework for sleep stage classification using single-channel electroencephalograms. *IEEE Trans Neural Syst Rehabil Eng*. 2022;**30**:2088–2096. doi: 10.1109/TNSRE.2022.3192988
38. Zhang H, Wang X, Li H, Mehendale S, Guan Y. Auto-annotating sleep stages based on polysomnographic data. *Patterns*. 2022;**3**(1):100371. doi: 10.1016/j.patter.2021.100371
39. Malafeev A, Laptev D, Bauer S, *et al*. Automatic human sleep stage scoring using deep neural networks. *Front Neurosci*. 2018;**12**:781. doi: 10.3389/fnins.2018.00781
40. Kingma DP, Ba JL. Adam: a method for stochastic optimization. In: proceedings of the 3rd International Conference on Learning Representations (ICLR); May 7, 2015; US: ICLR; pp.1–15.
41. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J Royal Stat Soc Series B: Stat Methodol*. 2005;**67**(2):768–768. doi: 10.1111/j.1467-9868.2005.00527.x
42. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. *Statist. Surv*. 2010;**4**:40–79.
43. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI*. 1995;**2**:1137.
44. Goldstein CA, Berry RB, Kent DT, *et al*. Artificial intelligence in sleep medicine: background and implications for clinicians. *J Clin Sleep Med*. 2020;**16**(4):609–618. doi: 10.5664/jcsm.8388