Benefits of Different Strategies to Adapt Sleep Scoring Models from Scalp- to Ear-EEG

Patrycja Lebiecka-Johansen, Jesper Strøm, Kaare Mikkelsen, Alvaro F. Cabrera, *Member, IEEE*, Rasmus E. Madsen, Julie A. E. Christensen, Martin C. Hemmsen and Preben Kidmose, *Member, IEEE*

***Abstract* — Polysomnography is an extensive evaluation of several physiological measures and the gold standard technique for clinical sleep assessment. However, this technique is both resource-expensive and often unfeasible over multiple nights. Ongoing research has shown that ear-EEG technology combined with the model-based automatic sleep scoring can be used for long-term sleep monitoring. More work is needed to robustly adapt clinical polysomnography-based sleep scoring models to at-home ear-EEG-based sleep patterns. Here, we investigated the main and combined benefits of utilizing three different strategies to adapt sleep scoring models from scalp- (part of polysomnography) to ear-EEG: 1) *fine-tuning of the sleep scoring model to left-right mastoid scalp-EEG, 2) fine-tuning of the sleep scoring model to ear-EEG* and *3) ensemble prediction.* The results showed that all strategies applied in isolation improve the sleep scoring performance on ear-EEG data relative to the not adapted model. With combined two or three strategies, sleep scoring performance on ear-EEG reaches performance comparable with sleep scoring on scalp-EEG (from κ 0.71 to 0.83; from κ 0.67 to 0.77; from κ 0.57 to 0.68 in three data sets).**

***Clinical Relevance*— This study demonstrates that current crosshead ear-EEG technology, combined with advanced sleep scoring models, enables accurate at-home sleep monitoring. Furthermore, tailoring sleep scoring models to ear-EEG data enhances identification of the sleep architecture.**

Keywords—ear-EEG, automatic sleep scoring, fine-tuning, left-right mastoid, ensemble prediction, U-Sleep

I. Introduction

Polysomnography (PSG) is a multimodal monitoring technique and the reference standard for clinical sleep assessment. PSG recording setups are technically complex and require supervision of clinical personnel, typically over a single or a few nights. PSG-based sleep assessments have several additional limitations. First, the intrusive nature of the recording setup affects the sleep patterns [1][2]. Second, single-night recordings do not account for inter-night variability, limiting the assessment of sleep dynamics over time. For these reasons, PSG is not feasible for long-term sleep monitoring.

Therefore, to facilitate more accessible and representative methods for sleep assessment, minimally obtrusive devices suitable for long-term sleep monitoring in the patient's natural environment are required. In pursuit of this goal, numerous portable and less intrusive devices have been proposed [3][4]. Among these emerging technologies for sleep assessment are so-called ear-EEG devices, which are based on electrodes around- or in-the-ear [5]. Furthermore, to fully leverage the potential of long-term sleep monitoring using portable devices - which typically feature a highly reduced set of sensing modalities - automated sleep assessment procedures are essential.

The objective of this study was to investigate various strategies for adapting sleep models from scalp-EEG (part of the PSG setup) to ear-EEG. In machine learning terminology, model adaptation constitutes a domain transfer learning problem and implicitly raises the question of differences between the two domains. Two main factors contribute to these differences. First, scalp-EEG and ear-EEG represent different projections of the underlying neural sources. More specifically, the lead field matrix, representing the transfer function from source space to measurement space, differs between the two modalities [6]. Second, the instrumentation and measurement setups for scalp-EEG and ear-EEG vary in several respects, including electrode type and size, wet or dry electrode-skin interfaces, methods for retaining electrodes against the body, and amplifier electrical characteristics [7]. These factors collectively affect noise properties and susceptibility to electrical, physiological, and motion artifacts.

To adapt sleep scoring models from scalp-EEG to ear-EEG, we assessed the potential benefits of three strategies, in isolation as well as in combination. Due to factors described above and previous work on optimal electrode configurations [8], we applied sleep scoring models to signal derivatives calculated across head by re-referencing the recorded signals. Those were left-right mastoid scalp-EEG derivative in PSG and crosshead ear-EEG derivative. Firstly, we fine-tuned a

* This project has been funded by the Innovation Fund Denmark, grant 9066-00021B.

P. Lebiecka-Johansen was with T&W Engineering, Allerød, Denmark and is with Center for Ear-EEG, Department of Electrical and Computer Engineering, Aarhus University, Aarhus, (phone: +45 50210811; e-mail: patrycja@ece.au.dk).

J. Strøm, K. Mikkelsen and P. Kidmose are with Center for Ear-EEG, Department of Electrical and Computer Engineering, Aarhus University Aarhus, Denmark (e-mails: js@ece.au.dk, mikkelsen.kaare@ece.au.dk, pki@ece.au.dk).

A. F. Cabrera, R. E. Madsen, J. A. E. Christensen and M. C. Hemmsen are with T&W Engineering, Allerød, Denmark (e-mails: alca@tweng.com, rem@tweng.com, jaec@tweng.com, mrhe@tweng.com)

Manuscript 889 submitted to 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) . Received February 7, 2025.

scalp-EEG-based sleep scoring model to a signal mimicking a crosshead ear-EEG derivative before testing it on the actual ear-EEG data. Secondly, we fine-tuned a scalp-EEG-based sleep scoring model to a relatively small subset of ear-EEG data and tested its performance on the unseen recordings [9]. Thirdly, we used multiple crosshead ear-EEG derivatives to predict sleep stages with a scalp-EEG-based sleep scoring model. This approach was built on the assumption that multiple crosshead ear-EEG would differ more in their measurement noise characteristics relative to their representation of the underlying neural activity. As a result, a combination in the form of ensemble learning would make model-based sleep scoring on ear-EEG more robust.

II. METHODS

A. Ear-EEG Data Sets

We used three ear-EEG data sets in the current study. Ear-EEG 1 data set was acquired and published in [5]. It contains recordings from 20 healthy younger subjects (μ = 25.9 years, range = [22, 36]) across 4 nights. Ear-EEG 2 data set was recorded and first published in [10]. It contains recordings from 10 healthy younger subjects (μ = 27.4 years, range = [22, 35]) across 2 nights. Ear-EEG 3 data set contains recordings from 15 healthy older (μ = 61.3 years, range = [57, 75]) as well as 7 healthy younger subjects (μ = 30.7 years, range = [25, 33]) from a single night. This data set is neither publicly available nor had it been published in a journal before. The data sets differ in the ear-EEG setup they were recorded with, which are described in Table I.

TABLE I. SUMMARY OF THE EAR-EEG SETUPS ACROSS DATA SETS.

	Ear-EEG 1	***Ear-EEG 2***	***Ear-EEG 3***
Subjects (Older O, Younger Y)	20 Y: μ = 25.9y, [22, 36]	10 Y: μ = 27.4y, [22, 35]	15 O: μ = 61.3y, [57,75] 7 Y: μ = 30.7y, [25, 33]
Recordings per subject	4	2	1
Ear-mold design	Custom	Generic	Custom
Electrode setup	6 left + 6 right	3 left + 2 right	3 left + 2 right
Reference	Average	Left ear	Left ear
Sampling frequency	500Hz	250Hz	250Hz

All ear-EEG recordings were acquired concurrently with PSG. For dataset 1, both PSG and ear-EEG recordings were recorded with the same amplifier, and therefore automatically aligned [5]. For datasets 2 and 3, the ear-EEG recordings were recorded with two different independent amplifiers and aligned with PSG based on physiological artifacts which were deliberately introduced at the beginning of the recordings [10]. Following the signal alignment, human experts' sleep stage labels on PSG recordings were used to label the corresponding epochs in the ear-EEG recordings [5], [10].

B. Pre-processing of the Ear-EEG

All data sets were pre-processed in the same way. To start with, the recorded ear-EEG signals were corrected for the DC-values by subtracting the mean, and bandpass filtered between 0.1 and 100Hz. Afterwards, notch filters at 50Hz and 100Hz were applied to remove a powerline noise. Then, artifacts including out-of-range values, electrode spikes, excessive subject movement and EMG artifacts were removed from the signal by an automated process described in [5]. Gaps caused by the artifacts' removal were set to empty values (NaNs). At last, recorded ear-EEG signals were resampled to the same sampling frequency 128Hz to align with the procedure used for training the original sleep scoring model as described in [11]. Recorded signals were divided into 30-s-long non-overlapping epochs and each epoch was associated with a label. Recorded signals consisting of more than 30% artifact-full samples were considered invalid, set to empty values (NaNs) and excluded from further analysis. A total of 134 recorded ear-EEG signals out of 1128 (11.9%) were considered invalid. Invalid signals were neither used to obtain crosshead ear-EEG derivatives nor to predict sleep stages.

To obtain crosshead ear-EEG derivatives, several steps were performed. First, recorded signals were re-referenced to the average of all signals (samples and derivatives previously considered as artefacts were not used to calculate the average). In the ear-EEG 2&3 data sets, the reference electrode (see Table I) was introduced as a vector filled with zeros of the same length as the recorded signals prior to the re-referencing. Secondly, a set of crosshead derivatives was calculated: 1) a single crosshead derivative was calculated by subtracting the average of the left ear from the average of the right ear [5]; 2) an average of the opposite ear was subtracted from each individual re-referenced signal. The second approach was used to predict sleep stages based on the ensemble of crosshead derivatives (see section II, E.2). Newly created crosshead ear-EEG derivatives constructed from invalid signals (see paragraph above) were removed from further analysis. At last, gaps were filled in with a temporal linear interpolation. The procedure ensured that all crosshead derivatives used for sleep scoring were comparable (harmonized) across datasets.

C. Sleep Scoring Labels

All ear-EEG recordings were acquired concurrently with PSG. Professional sleep scorers assigned labels to all 30-s-long epochs in the PSG recordings following the American Academy of Sleep Medicine Manual [12]. Sleep scorers were blinded to the ear-EEG data. Epochs that were difficult to score by a sleep scorer were marked as UNKNOWN. Furthermore, epochs containing any invalid samples (see section II B) were re-labelled as NOISE. This was done for every crosshead ear-EEG derivative separately.

Repeated re-labelling process created redundancy in labels across epochs (NOISE or a sleep stage). To reduce the number of labels per epoch (across crosshead derivatives) to one, the rule of majority voting was used – the label of a given epoch was chosen as the most popular one across crosshead derivatives. In case of several labels being equally popular, the first-occurring most voted label was chosen. On average 11,8% of labels per data set were marked as NOISE (excluding invalid crosshead derivatives). Epochs labelled as UNKNOWN and NOISE were excluded from the calculation of the validation metrics during fine-tuning and test metrics

Manuscript 889 submitted to 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) . Received February 7, 2025.

during evaluation of sleep scoring models. Eight nights out of 122 (6.6%) were considered invalid since either majority of recorded ear-EEG signals had poor quality, or majority of crosshead ear-EEG derivatives was re-labelled as NOISE. Those nights were not used in the further analysis.

D. Sleep Scoring Model

1) U-Sleep: A fully convolutional deep learning model U-Sleep was used as a sleep scoring model [11]. U-Sleep was trained on a large pool of PSG recordings representing both healthy and pathological populations, ethnicities, biological sexes as well as diverse age groups. It showed excellent performance, closely matching human expert scorings, across data sets. U-Sleep was originally trained using a set of a single random scalp-EEG channel (that is EEG signal recorded from a pre-defined electrode position) and a single EOG channel, both being a part of the clinical PSG setup. Implementation details can be found in the Supplementary Material of [11].

In the manual sleep scoring, EOG is specifically important for identifying REM sleep. However, this study focused on sleep scoring on ear-EEG, which does not provide a dedicated EOG channel. For this reason, we decided to train the U-Sleep model as a single channel model based on a randomly selected scalp-EEG channel. We used our own pyTorch implementation of the U-Sleep pipeline described in [13].

2) Performance Metric: To evaluate the sleep scoring model performance, we used a metric called Cohen's κ [14]. Cohen's κ measures inter-rater agreement, and accounts for by-change agreement. Thus, it is suitable for quantifying the agreement between model sleep stage predictions and human expert labels [5].

E. Three Strategies to Adapt Sleep Scoring Models from Scalp- to Ear-EEG

1) Fine-tuning to Left-Right Mastoid: To adapt the model to the ear-EEG data without exposing it to the actual data, the U-Sleep model was fine-tuned to a pool of left-right mastoid scalp-EEG derivatives in a subset of the PSG recordings. Two conditions were specified to select a satisfactory subset: 1) electrodes placed on both right and left mastoids must have been included in the PSG electrode setup; 2) there existed a common for both mastoids reference in the scalp-EEG. 7570 out of 19359 available PSG recordings fulfilled the requirements. The training pipeline for both the single-channel-based U-Sleep and fine-tuning to left-right mastoid was identical to the one described in [13].

2) Fine-tuning to Ear-EEG: Here, the U-Sleep pre-trained on scalp-EEG was fine-tuned to a combined ear-EEG 1 and 2 data set consisting of a total of 100 nights. We decided to leave ear-EEG 3 out of the fine-tuning process due to 1) its smaller size with a single night per person; 2) presence of the two age groups in this data set. As ear-EEG 1 and 2 but not ear-EEG 3 were used for fine-tuning the U-Sleep model, two procedures were used to evaluate the sleep scoring performance.

In ear-EEG 1 and 2 sleep scoring performance was evaluated through a cross-validation technique. That is, 30 subject-models were fine-tuned, where the subject under evaluation was not seen by the model during training and validation. To do that, for each subject-model, the combined ear-EEG 1 and 2 data set was divided into a 1) training set with 26 randomly selected subjects (78-80 nights per loop), 2) validation set with 3 subjects selected in a stratified way such that 2 are from the larger ear-EEG 1 set and 1 from the smaller ear-EEG 2 set (18 nights per loop), 3) test set with a single left-out subject (2 or 4 nights). Fine-tuning was performed with 50 training + validation iterations for every subject-model. A single training iteration consisted of 5 mini-batches, that is 5 batches consisting of 64 (batch size) segments of 35x30s-long sleep epochs, which were obtained with a semi-random sampling technique [13]. The number of subjects in the validation set and number of mini-batches per training epoch were selected based on a grid search. In this grid search, training was ran with different numbers of subjects in the validation data set {2, 3, 4, 5} and mini-batches per training iteration {5, 10, 15, 20}. A pair of optimal parameters was selected based on the largest obtained validation Cohen's κ and lowest variability across validation subjects. Network architecture and architecture-specific parameters were kept identical to the original U-Sleep training pipeline, except for the number of input channels (single randomly selected crosshead ear-EEG derivative) [11],[13]. Sleep scoring performance on the ear-EEG 1 and 2 data sets was assessed in a testing stage, where test Cohen's κ was calculated for the unseen subject in every best in terms of validation Cohen's κ out of 50 iterations subject-model.

To assess sleep scoring performance on the ear-EEG 3 data set, a U-Sleep model was fine-tuned to a combined ear-EEG 1 and 2 data set. This time, ear-EEG 1 and 2 data sets were divided into training and validation sets (see above), without a separate test set, and performed up to 50 training iterations with a single split and optimal parameters. The best fine-tuned model was selected based on the validation Cohen's κ with the early stopping criterion – 3 consecutive iterations without increase in Cohen's κ by min. 0.01 terminated the fine-tuning. Sleep scoring performance on ear-EEG 3 data set was assessed after finalizing the fine-tuning process.

3) Ensemble Prediction: Here, multiple crosshead ear-EEG derivatives (see section II B) were used to predict sleep stages. Figure 1 presents the ensemble prediction.

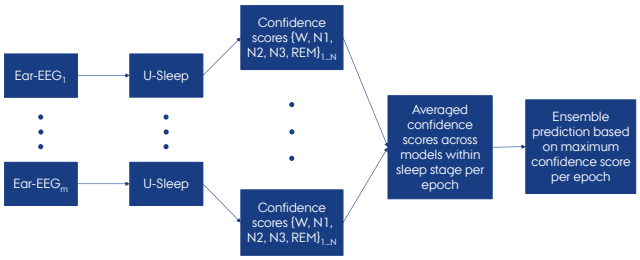

Figure 1. Scheme of the ensemble prediction as a strategy to adapt sleep scoring model from scalp- to the ear-EEG data.

To start with, every crosshead ear-EEG derivative (12 or 5 depending on the data set) was passed through the same version of the U-Sleep model, independently of each other. We then obtained confidence scores for all sleep stages (W, N1, N2, N3, REM), across available epochs for each crosshead ear-EEG derivative. To obtain an ensemble average, confidence scores were averaged within each sleep stage, across all valid crosshead ear-EEG derivatives for every valid epoch (that is within-derivative-epochs not considered invalid, see section II B). Lastly, sleep stage with the highest confidence score per epoch was used as a prediction. The process of obtaining ensemble prediction vector (EP) is summarized in (1).

$$EP = \left[max\left\{ \left\{ {}^{1}/_{|S_n \notin \emptyset|} \sum_{m=1}^{M} P(s_{mn} = C_i) \right\}_{i \in \{W,N1,N2,N3,REM\}} \right\}^{(n)} \right]_1^N \quad (1)$$

EP is a 1x N-epochs-long vector of sleep stage predictions ($C_i \in \{W,\ N1, N2, N3, REM\}$). The sleep stage prediction in each epoch n was based on the maximum confidence score ($P(S_{mn} = C_i)$) among the averaged across non-empty crosshead ear-EEG derivatives, denoted as S_n. The number of non-empty crosshead ear-EEG derivatives denoted as M ($M = S_n \notin \emptyset$) was dynamically calculated in every epoch.

F. Statistical Analysis

1) Linear Mixed Modelling: Linear mixed modelling (LME) framework was used to assess the benefits of the three tested strategies in isolation and interaction. The advantage of LME is that it provides flexibility in including both fixed factors, that are factors controlled by the experimental design, and random factors, that are factors influencing variance in the data not controlled by the design. Furthermore, it can handle unbalanced data sets (such as different numbers of nights and subjects within a data set) and allows for the analysis of effect sizes. LME was implemented with the lme4 package in RStudio [15]. To assess the effects of applied strategies in a full factorial experimental design, an LME was fitted to test Cohen's κ from the combined ear-EEG 1,2 and 3 data sets.

$$Cohen's\ \kappa = FT_{LRmastoid} * Ensemble * FT_{earEEG} + (1|subject:dataset) + (1|night:dataset) + (1|age\ group) \quad (2)$$

The first part of the right-hand side equation implies that both main and 2- as well as 3-way interactions were investigated. The latter implies that a combination of nested and cross factors was included. The interindividual and repeated measures effects were estimated within data sets (nested factors). Age group was added as a non-nested random factor, since it related to a single data set. Akaike's Information Criterion and Log Likelihood values were used to optimize the model fit, so that it included all relevant random factors, reached convergence and did not produce singularities in the estimated random effects.

2) Estimated Fixed Effects and Analysis of Contrasts: LME-ANOVA was used to address the hypotheses, that is to test for main and interaction effects in the fitted LME model. Then, a contrast analysis was performed on the estimated marginal means to better understand the direction and size of effects with the 'emmeans' package in RStudio [16]. Specifically, pairwise t-tests were performed to identify the magnitude of performance improvement due to the applied strategies. In addition, the effect sizes of pairwise comparisons were investigated to understand the scale of the improvements based on the estimated confidence intervals [17]. The effect sizes were calculated with Cohen's d. Pairwise comparisons were performed on each contrast within a fixed factor with the two other factors frozen. More specifically, the following contrasts were tested: $FT_{LRmastoid}$: Yes vs. No, Ensemble: Yes vs. No, and $FT_{ear\text{-}EEG}$: Yes vs. No.

III. Results

A. Cohen's κ Across Tested Conditions

Fig. 2 shows κ medians and interquartile ranges across tested conditions. Labels indicate which strategy has been applied to adapt the scalp-EEG-based U-Sleep to ear-EEG model. Descriptive statistics on raw test Cohen's κ are summarized in Table II. Best performance for ear-EEG 1 data set was observed with the U-Sleep fine-tuned to ear-EEG using an ensemble of crosshead ear-EEG derivatives for predictions ($FT_{ear\text{-}EEG}$ & Ensemble, μ = 0.81, sd = 0.06 and η = 0.83, IQR = 0.06). In ear-EEG 2&3, a combination of all three adaptations yielded the best performance ($FT_{LRmastoid}$ & $FT_{ear\text{-}EEG}$ & Ensemble; ear-EEG data set 2 μ = 0.73, sd = 0.12; ear-EEG 3 μ = 0.64, sd = 0.14 and η = 0.68, IQR = 0.18).

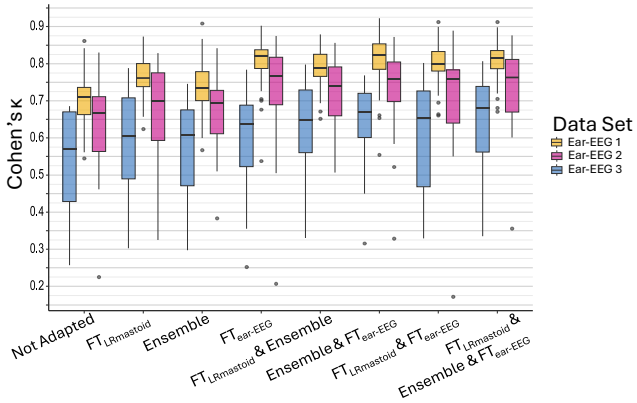

Figure 2. The benefits of strategies for adapting sleep scoring model from scalp- to the ear-EEG data. Median η of model performance Cohen's κ across tested conditions. Labels represent the applied strategies. Black dots represent nights outside of the 1.5*IQR. FT – Fine-tuning. $FT_{LRmastoid}$ adaptation: U-Sleep pre-trained on LRmastoid scalp-EEG derivative; Ensemble adaptation: an ensemble of crosshead ear-EEG derivatives used to predict sleep stages; $FT_{ear\text{-}EEG}$ adaptation: U-Sleep fine-tuned to the ear-EEG data.

Manuscript 889 submitted to 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) . Received February 7, 2025.

Lowest median sleep scoring performance was observed for the not adapted U-Sleep applied to ear-EEG data (1) Not adapted, ear-EEG 1 $\eta = 0.71$, IQR = 0.08; ear-EEG 2 $\eta = 0.67$, IQR = 0.15; ear-EEG 3 $\eta = 0.57$, IQR = 0.24). To assess the benefits of applied strategies independent of the non-controllable noise sources, differences in sleep scoring performance were analyzed with Linear Mixed Models.

TABLE II. DESCRIPTIVE STATISTICS ON RAW K ACROSS CONDITIONS.

Condition	*Ear-EEG*	*Mean μ (sd)*	*Median η (IQR)*
Not adapted	1	0.70 (0.06)	0.71 (0.08)
	2	0.63 (0.14)	0.67 (0.15)
	3	0.53 (0.15)	0.57 (0.24)
Fine-tuning$_{LRmastoid}$	1	0.76 (0.05)	0.76 (0.06)
	2	0.67 (0.12)	0.70 (0.25)
	3	0.60 (0.14)	0.61 (0.22)
Ensemble	1	0.74 (0.06)	0.74 (0.08)
	2	0.67 (0.11)	0.70 (0.12)
	3	0.58 (0.13)	0.61 (0.21)
$FT_{ear-EEG}$	1	0.81 (0.06)	0.82 (0.05)
	2	0.72 (0.16)	**0.77 (0.13)**
	3	0.60 (0.14)	0.64 (0.17)
$FT_{LRmastoid}$ & Ensemble	1	0.79 (0.05)	0.79 (0.06)
	2	0.72 (0.10)	0.74 (0.13)
	3	0.63 (0.13)	0.65 (0.17)
Ensemble & $FT_{ear-EEG}$	**1**	**0.81 (0.06)**	**0.83 (0.06)**
	2	0.73 (0.13)	0.76 (0.11)
	3	0.63 (0.13)	0.67 (0.12)
$FT_{LRmastoid}$ & $FT_{ear-EEG}$	1	0.80 (0.05)	0.80 (0.05)
	2	0.71 (0.15)	0.76 (0.14)
	3	0.62 (0.14)	0.65 (0.26)
$FT_{LRmastoid}$ & Ensemble & $FT_{ear-EEG}$	1	0.81 (0.05)	0.82 (0.05)
	2	**0.73 (0.12)**	0.76 (0.14)
	3	**0.64 (0.14)**	**0.68 (0.18)**

Condition: FT: Fine-tuning. LRmastoid: U-Sleep pre-trained on LRmastoid scalp-EEG derivative; Ensemble: an ensemble of crosshead ear-EEG derivatives used to predict sleep stages; ear-EEG: U-Sleep fine-tuned to the ear-EEG data

Numbers in bold indicate the best obtained sleep scoring performance per data set.

B. Random Effects of the Linear Mixed Model

Fitted full-factorial model explained 6.6% of marginal variance and 89% of conditional variance in the data (R^2). The estimated random effects, that is variance explained by random factors, were as follows: $\sigma^2_{subject:data\_set} = 0.01$, sd = 0.1, $\sigma^2_{age\_group} = 0.006$, sd = 0.08, $\sigma^2_{night:data\_set} = 0.00$, sd = 0.02. Residual, unexplained variance was estimated as $\sigma^2 = 0.002$, sd = 0.05.

C. Mean and Interaction Effects

To address the hypotheses of the study, LME-ANOVA was performed on the estimated fixed effects. Table III shows results of this test. As hypothesized, all strategies to adapt the sleep scoring model to ear-EEG improved its performance in isolation ($FT_{LRmastoid}$: $F_{(1, 881.64)} = 77.56$, $p < 0.001$; Ensemble: $F_{(1, 881.64)} = 48.14$, $p < 0.001$; $FT_{ear-EEG}$: $F_{(1, 881.64)} = 319.89$, $p < 0.001$). Furthermore, the LME-ANOVA test revealed an interaction effect $FT_{LRmastoid}$ x $FT_{ear-EEG}$ ($F_{(1, 881.64)} = 82.76$, $p < 0.001$) and Ensemble x $FT_{ear-EEG}$ ($F_{(1, 881.64)} = 12.58$, $p < 0.001$). To be able to rank the applied strategies by their impact, a pairwise comparisons' analysis of the estimated marginal means was performed, and Cohen's d effect sizes were calculated.

TABLE III. MAIN AND INTERACTION EFFECTS OF THE ASSESSED STRATEGIES THE SLEEP SCORING MODEL'S PERFORMANCE ON THE EAR-EEG DATA.

	Sum of Squares	*F(df), p-value*
$FT_{LRmastoid}$	0.167	$F_{(1, 881.64)} = 77.56$, **$p < 0.001$**
Ensemble	0.104	$F_{(1, 881.64)} = 48.14$, **$p < 0.001$**
$FT_{ear-EEG}$	0.69	$F_{(1, 881.64)} = 319.89$, **$p < 0.001$**
$FT_{LRmastoid}$ & Ensemble	0.00	$F_{(1, 881.64)} = 0.02$, $p = 0.89$
$FT_{LRmastoid}$ & $FT_{ear-EEG}$	0.178	$F_{(1, 881.64)} = 82.76$, **$p < 0.001$**
Ensemble & $FT_{ear-EEG}$	0.027	$F_{(1, 881.64)} = 12.58$, **$p < 0.001$**
$FT_{LRmastoid}$ & $FT_{ear-EEG}$ & Ensemble	0.002	$F_{(1, 881.64)} = 0.92$, $p = 0.34$

D. Estimated Marginal Means and Pairwise Comparisons

Table IV shows the result of pairwise comparisons on the estimated marginal means. Fine-tuning U-Sleep model to the crosshead ear-EEG data significantly improved the sleep scoring performance in all tested pairwise comparisons (9) Δ 0.096, $p < 0.001$; 10) Δ 0.035, $p < 0.001$; 11) Δ 0.068, $p < 0.001$; 12) Δ 0.019, $p < 0.001$). The benefit of fine-tuning U-Sleep to the left-right mastoid scalp-EEG data was observed in absence of fine-tuning to crosshead ear-EEG (1) Δ 0.058, $p < 0.001$; 2) Δ 0.035, $p < 0.001$). Similarly, a benefit to sleep scoring performance was observed with the ensemble prediction (5) Δ 0.035, $p < 0.001$; 6) Δ 0.029, $p < 0.001$).

In terms of Cohen's d effect sizes, the largest improvement in sleep scoring performance was observed when fine-tuning U-Sleep to ear-EEG in isolation (9) d = 2.06, CL = [1.78, 2.33]). Very large (1.3 <) or large (0.8-1.3) improvement to sleep scoring performance in terms of effect size [13] was also observed for fine-tuning U-Sleep to crosshead ear-EEG in combination with ensemble prediction (11) d = 1.47, CL = [1.2, 1.74]) and fine-tuning to left-right mastoid scalp-EEG without ensemble (1) d = 1.24, CL = [0.98, 1.51] as well as with ensemble (2) d = 1.1, CL = [0.84, 1.36]). Ensemble prediction had a moderate improvement of sleep scoring performance yet only in absence of fine-tuning to ear-EEG (5) d = 0.76, CL = [0.5, 1.02]; d = 0.61, CL = [0.35, 0.87]).

TABLE IV. PAIRWISE COMPARISONS – CONTRASTS ADJUSTED FOR MULTIPLE COMPARISONS AND COHEN'S D EFFECT SIZES

Frozen Factors (Conditions)[a]	**Statistics**		
	Contrast	*Estimated Marginal Difference[b]*	*Effect size (Cohen's d)[c]*
Ensemble = No **$FT_{ear-EEG}$ = No**	**1) $FT_{LRmastoid}$: Yes – No**	**Δ 0.058, $t_{(889)} = 9.46$, $p < 0.001$**	**d = 1.24, CL = [0.98, 1.51]**
Ensemble = Yes **$FT_{ear-EEG}$ = No**	**2) $FT_{LRmastoid}$: Yes – No**	**Δ 0.051, $t_{(889)} = 8.37$, $p < 0.001$**	**d = 1.1, CL = [0.84, 1.36]**
Ensemble = No $FT_{ear-EEG}$ = Yes	3) $FT_{LRmastoid}$: **Yes – No**	Δ -0.003, $t_{(889)} = -0.55$, p = n. s.	d = -0.07, CL = [-0.33, 0.19]
Ensemble = Yes $FT_{ear-EEG}$ = Yes	4) $FT_{LRmastoid}$: **Yes – No**	Δ 0.001, $t_{(889)} = -0.26$, p = n. s.	d = 0.03, CL = [-0.22, 0.29]
$FT_{LRmastoid}$ = No **$FT_{ear-EEG}$ = No**	**5) Ensemble: Yes – No**	**Δ 0.035, $t_{(889)} = 5.77$, $p < 0.001$**	d = 0.76, CL = [0.5, 1.02]

Frozen Factors (Conditions)[a]	Statistics		
	Contrast	*Estimated Marginal Difference*[b]	*Effect size (Cohen's d)*[c]
$FT_{LRmastoid}$ = Yes **$FT_{ear-EEG}$ = No**	**6) Ensemble: Yes – No**	**Δ 0.029, $t_{(889)}$ = 4.68, p < 0.001**	d = 0.61, CL = [0.35, 0.87]
$FT_{LRmastoid}$ = No $FT_{ear-EEG}$ = Yes	7) Ensemble: **Yes – No**	Δ 0.007, $t_{(889)}$ = 1.28, p = n. s.	d = 0.17, CL = [-0.09, 0.43]
$FT_{LRmastoid}$ = Yes $FT_{ear-EEG}$ = Yes	8) Ensemble: **Yes – No**	Δ 0.013, $t_{(889)}$ = 2.1, p = n. s.	d = 0.28, CL = [0.02, 0.53]
$FT_{LRmastoid}$ = No **Ensemble = No**	**9) $FT_{ear-EEG}$: Yes - No**	**Δ 0.096, $t_{(889)}$ = 15.68, p < 0.001**	**d = 2.06, CL = [1.78, 2.33]**
$FT_{LRmastoid}$ = Yes **Ensemble = No**	**10) $FT_{ear-EEG}$: Yes – No**	**Δ 0.035, $t_{(889)}$ = 5.67, p < 0.001**	d = 0.74, CL = [0.48, 1.00]
$FT_{LRmastoid}$ = No **Ensemble = Yes**	**11) $FT_{ear-EEG}$: Yes – No**	**Δ 0.068, $t_{(889)}$ = 11.19, p < 0.001**	**d = 1.47, CL = [1.2, 1.74]**
$FT_{LRmastoid}$ = Yes **Ensemble = Yes**	**12) $FT_{ear-EEG}$: Yes - No**	**Δ 0.019, $t_{(889)}$ = 3.09, p = 0.02**	d = 0.41, CL = [0.15, 0.66]

[a]Conditions: FT – Fine-tuning. LRmastoid: U-Sleep pre-trained on LRmastoid scalp-EEG derivative; Ensemble: an ensemble of crosshead ear-EEG derivatives used to predict sleep stages; ear-EEG: U-Sleep fine-tuned to the ear-EEG data

[b]Significance threshold p = 0.05, significant differences in bold

[c]0.2-0.5 small; 0.5-0.8 medium; **0.8-1.3 large; 1.3 < very large effect size**

IV. Discussion

This study aimed to investigate strategies for adapting sleep scoring models from scalp- to ear-EEG data. We investigated the main and interaction effects of utilizing three strategies: *1) fine-tuning of the sleep scoring model to left-right mastoid scalp-EEG, 2) fine-tuning of the sleep scoring model to crosshead ear-EEG* and *3) ensemble prediction*. The results showed that all three strategies in isolation improved the sleep scoring model performance on ear-EEG data relative to the not adapted model. Fine-tuning the sleep scoring model to ear-EEG data was found especially beneficial in terms of effect size. Combining fine-tuning to ear-EEG and ensemble prediction (and sometimes fine-tuning to left-right mastoid scalp-EEG derivative) showed largest improvement to sleep scoring performance (uncorrected for inter- and intra-differences).

A. The Benefits of Strategies to Adapt Sleep Scoring Models on Their Performance

Both Cohen's κ and statistical analyses show that it is advantageous to adapt sleep scoring models from scalp- (PSG) to ear-EEG data. Furthermore, it seems that choosing the right strategy matters for the results. In this study we found that fine-tuning the U-Sleep model to the ear-EEG data yields generally the largest improvement. Since some sleep patterns may be represented differently at the scalp domain as compared to the ear domain, this result was expected.

Interestingly, fine-tuning a sleep scoring model to left-right mastoid scalp-EEG derivative has almost as large positive effect on the performance as fine-tuning it directly to ear-EEG (in terms of the effect size). This result aligns with the literature [6] and demonstrates that the left-right mastoid derivative obtained from scalp-EEG and the crosshead derivative obtained from ear-EEG reflect similar cortical activity.

To our surprise, ensemble prediction was found less effective than the other two strategies applied in isolation (only a moderate benefit on sleep scoring performance was observed). However, a combination of ensemble prediction and the fine-tuning approaches showed largest benefits to sleep scoring performance across data sets. While the ensemble prediction may not be needed all the time, it may be an effective way to clean the data and to increase robustness of the scoring. For example, ensemble prediction could be used as a dynamic quality-based re-weighting of the crosshead ear-EEG derivatives to predict sleep stages exclusively from the good quality data. More work is needed to define the optimal method for such a quality-based re-weighting.

B. Limitations

In the current study, we decided to evaluate the benefits of applied strategies to adapt sleep scoring from scalp- to ear-EEG by means of a model performance across harmonized data sets. Despite of a systematic harmonization of the data sets, some differences could have contributed to the results in an unaccounted way. For example, ear-EEG 2 and 3 were smaller in size (number of subjects and nights per subject) than ear-EEG 1. Furthermore, instrumentation and measurement setups as well as the number of recorded nights per subject varied across data sets. Finally, subject characteristics differed across data sets. Thus, we suggest that a search of optimal strategy should be extended with larger and more diversified ear-EEG data sets in the future.

V. Conclusions

This study shows that adapting a sleep scoring model from crosshead scalp- to ear-EEG data improves performance. A combination of fine-tuning a sleep scoring model to ear-EEG and predictions based on the ensemble of crosshead ear-EEG derivatives was found to be the most effective strategy.

Acknowledgment

Authors thank Nelly Shenton, Yousef R. Tabar, and Mia Hansen for their involvement in designing the original studies, recruiting participants and acquiring data. Special thanks to Marit Otto and Mia D. Thomsen for labeling the PSG recordings. Authors thank Lars D. Mosgaard for creating a hardware cleaning pipeline. Besides, authors want to thank all the study participants for their willingness to test the ear-EEG technology throughout multiple nights.

References

[1] O. Le Bon, L. Staner, G. Hoffmann, M. Dramaix, I. S. Sebastian, J.R. Murphy, M. Kentos, I. Pelc and P. Linkowski, The first-night effect may last more than one night,, *Journal of Psychiatric Research*,, vol. 35, 3, 2001, pp. 165-172, doi:10.1016/S0022-3956(01)00019-X

[2] J. Newell, O. Mairesse, P. Verbanck, and D. Neu, "Is a one-night stay in the lab really enough to conclude? First-night effect and night-to-night variability in polysomnographic recordings among different clinical population samples", *Psychiatry Research*, vol. 200, 2–3, 2012, pp. 795-801, doi:10.1016/j.psychres.2012.07.045

[3] C.J. de Gans, P. Burger, E.S. van den Ende, J. Hermanides, P.W.B. Nanayakkara, R.J.B.J. Gemke, F. Rutters and D.J. Stenvers,. "Sleep assessment using EEG-based wearables–A systematic review." *Sleep Medicine Reviews*, p.101951, 2024. doi: 10.1016/j.smrv.2024.101951

[4] M., Mohamed, N. Mohamed, and J.G. Kim, "Advancements in Wearable EEG Technology for Improved Home-Based Sleep Monitoring and Assessment: A Review." *Biosensors*, 13(12), p.1019, 2023. doi: 10.3390/bios13121019

[5] K. Mikkelsen, S. L. Kappel, C. B. Christensen, H. O. Toft, M. C. Hemmsen, M. L. Tank, M. Otto and P. Kidmose, "Accurate whole-night sleep monitoring with dry-contact ear-EEG", *Scientific Reports*, vol. 9, 16824, 2019. doi: 10.1038/s41598-019-53115-3

[6] S. L. Kappel, S. Makeig and P. Kidmose, "Ear-EEG Forward Models: Improved Head-Models for Ear-EEG", *Frontiers in Neuroscience*, vol. 13, 2019, doi: 10.3389/fnins.2019.00943

[7] S.L. Kappel, M.L. Rank, N.O. Toft, M. Andersen and P. Kidmose. "Dry-contact electrode ear-EEG" *IEEE Trans. on Biomedical Engineering*, vol. 66, no. 1, pp. 150-158, Jan. 2019, doi: 10.1109/TBME.2018.2835778.

[8] K. B. Mikkelsen , H. Phan , M. L. Rank, M. C. Hemmsen, M. de Vos and P. Kidmose, "Sleep Monitoring Using Ear-Centered Setups: Investigating the Influence From Electrode Configurations", *IEEE Transactions On Biomedical Engineering,* vol. 69, no. 5, pp. 1564-1572, May 2022, doi: 10.1109/TBME.2021.3116274 2022

[9] K. P. Lorenzen, E. R. M. Heremans, M. de Vos and K. B. Mikkelsen, "Personalization of Automatic Sleep Scoring: How Best to Adapt Models to Personal Domains in Wearable EEG", *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 10, pp. 5804-5815, Oct. 2024, doi: 10.1109/JBHI.2024.3409165

[10] Y. R. Tabar, K. B. Mikkelsen, N. Shenton, S. L. Kappel, A. R. Bertelsen, R. Nikbakht, H. O. Toft, C. H. Henriksen, M. C. Hemmsen, M. L. Rank, M. Otto and P. Kidmose, "At-home sleep monitoring using generic ear-EEG", *Frontiers in Neuroscience*, vol. 17, Feb. 2023, doi: 10.3389/fnins.2023.987578

[11] M. Perslev, S. Darkner, L. Kempfner, M. Nikolic, P. J. Jennum and C. Igel, "U-Sleep: resilient high-frequency sleep staging", *npj Digit. Med.* vol. 4, 72, 2021, doi:10.1038/s41746-021-00440-5

[12] R. B. Berry, R. Brooks, C. E. Garnaldo, S. M. Harding, R. M. Lloyd, S. F. Quan, M. M. Troester and B. V. Vaughn, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications 2.5* (American Academy of Sleep Medicine, 2018)

[13] J. Strøm, A.L. Engholm, K.P. Lorenzen and K.B. Mikkelsen, "Common sleep data pipeline for combined data sets". *PLoS ONE*, vol. 19(8): e0307202. 2024, doi: 10.1371/journal.pone.0307202

[14] Cohen, J. A "Coefficient of Agreement for Nominal Scales". *Educational and Psychological Measurement*, vol. 20*,* 37–46 1960, doi: 10.1177/001316446002000104

[15] D. Bates, M. Mächler, B. Bolker and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4". *Journal of Statistical Software*, *67*(1), 1–48, 2015, doi: 10.18637/jss.v067.i01

[16] R. Lenth, "emmeans: Estimated Marginal Means, aka Least-Squares Means". R package version 1.10.5, https://rvlenth.github.io/emmeans/

[17] G. M. Sullivan and R. Feinn, "Using Effect Size—or Why the P Value Is Not Enough", *J Grad Med Educ.* vol Sep 4(3):279-82, 2012, doi: 10.4300/JGME-D-12-00156.1