Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

Expert-level sleep staging using an electrocardiography-only feed-forward neural network

Adam M. Jones[1*], Laurent Itti[1], and Bhavin R. Sheth[2,3]

[1]Neuroscience Graduate Program, University of Southern California, Los Angeles, California.
[2]Department of Electrical & Computer Engineering, University of Houston, Houston, Texas.
[3]Center for NeuroEngineering and Cognitive Systems, University of Houston, Houston, Texas.
[*]Corresponding author email: adamjones@alumni.usc.edu

Abstract

Reliable classification of sleep stages is crucial in sleep medicine and neuroscience research for providing valuable insights, diagnoses, and understanding of brain states. The current gold standard method for sleep stage classification is polysomnography (PSG). Unfortunately, PSG is an expensive and cumbersome process involving numerous electrodes, often conducted in an unfamiliar clinic and annotated by a professional. Although commercial devices like smartwatches track sleep, their performance is well below PSG. To address these disadvantages, we present a feed-forward neural network that achieves gold-standard levels of agreement using only a single lead of electrocardiography (ECG) data. Specifically, the median five-stage Cohen's kappa is 0.725 on a large, diverse dataset of 5 to 90-year-old subjects. Comparisons with a comprehensive meta-analysis of between-human inter-rater agreement confirm the non-inferior performance of our model. Finally, we developed a novel loss function to align the training objective with Cohen's kappa. Our method offers an inexpensive, automated, and convenient alternative for sleep stage classification—further enhanced by a real-time scoring option. Cardiosomnography, or a sleep study conducted with ECG only, could take expert-level sleep studies outside the confines of clinics and laboratories and into realistic settings. This advancement democratizes access to high-quality sleep studies, considerably enhancing the field of sleep medicine and neuroscience. It makes less-expensive, higher-quality studies accessible to a broader community, enabling improved sleep research and more personalized, accessible sleep-related healthcare interventions.

Keywords: sleep, stages, polysomnography, electrocardiography, cardiosomnography, deep learning

1. Introduction

Understanding sleep stages aids in the comprehension of many brain states and unconscious processes [1]. Stage classification was first formalized by Rechtschaffen and Kales (R&K) in 1968 [2] and later updated by the American Academy of Sleep Medicine (AASM) in 2007 [3]. This system categorizes sleep into Wake, rapid eye movement (REM), and Non-REM (NREM) stages 1 through 3 (N1, N2, N3). These stages consist of distinct brain activity and occur in a cyclic pattern. Furthermore, they are associated with specific physiological and neurological processes, such as waste clearing and particular types of memory consolidation [4], [5].

Traditionally, the gold standard for clinically relevant sleep staging, or sleep stage scoring, has been polysomnography (PSG). During PSG, an individual typically spends one or more nights in a clinic. While sleeping, they wear electrodes that collect at least a dozen channels of biophysical

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

data. These inputs include brain activity (electroencephalography, EEG), eye movement, muscle tension, heart activity (electrocardiography, ECG), and respiration. Afterward, an experienced human scorer annotates the night for stages and other pertinent events. Unfortunately, the inherent subjectivity among human sleep scorers leads to the lack of a definitive “ground truth”.

This subjectivity underscores the necessity for measuring performance in terms of inter-rater agreement, with Cohen’s kappa being the statistic of choice [6], [7]. Kappa quantifies the agreement between raters, adjusting for chance agreement, thus providing a more accurate reflection of genuine agreement than a simple percentage metric. To address the considerable expense, time requirements, and inconsistency associated with human annotation, automated methods for sleep stage classification are increasingly being developed and implemented [8], [9].

In the decades since the formalization of sleep research, there has been an increased understanding of the importance of sleep. This understanding, in turn, has fueled more sleep research, the healthcare community’s interest in sleep monitoring for precision medicine [10], and now the public’s burgeoning interest. However, there are two issues with PSG that make it a non-starter for widespread adoption. The first is the considerable cost of PSG in terms of human labor and equipment. The second is the sheer cumbersomeness for the sleeper (i.e., wearing electrodes, while non-invasive, is intrusive and inconvenient, leading, at least sometimes, to unrepresentative data).

The response to these issues has been the emergence of alternative methods that aim to reduce the cost and inconvenience by measuring other physiological signals [11]. Notably, methods using ECG offer advantages such as the need for fewer electrodes (three versus dozens) and a stronger signal than EEG. Despite the promising performance, there is still substantial room for improvement. Many studies and devices ignore harder-to-classify stages, exclude subjects based on demographic factors like age or health status, and show poor agreement with human-scored PSG. The ideal “sweet spot” would be a cheap, patient-friendly method that agrees with classical PSG.

This paper focuses on the potential of ECG because, as a raw signal, it is a powerful, non-invasive tool for monitoring the autonomic nervous system activity during sleep. The activity of the autonomic nervous system, like the central nervous system, noticeably changes during sleep. While often viewed as subordinate to the central nervous system, research has highlighted when autonomic activity precedes central activity, substantiating complex, bidirectional interactions [12]. Novel statistical approaches have further elucidated the dynamics of this flow during sleep, revealing activity patterns specific to each sleep stage [13].

In general, two issues have hampered efforts to score sleep stages automatically at an expert level without the full complement of biophysical inputs that PSG requires. First, researchers originally defined sleep stages primarily by their manifestations in EEG. The assumption that the brain input is, by definition, necessary has impeded the search for equally informative surrogates. The second obstacle has been the lack of enormous datasets for training generalizable classifiers. Today, tens of thousands of recordings of expert human-scored PSG are available for free, which makes it possible to overcome both issues. **Specifically, we set out to determine if it is possible to score sleep stages as well as human-scored PSG using only a single lead of ECG data.**

Here, we demonstrate a neural network for sleep staging that achieves gold-standard levels of agreement with PSG using only ECG data, forgoing traditional PSG inputs like brain, eye, and muscle activity. We trained the model on 4,000 recordings from subjects 5 to 90 years old to improve the generalizability. Furthermore, we developed a new loss function to align the training

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

objective with kappa. In addition to the excellent performance, we show that the model is robust and concordant with human-scored PSG. Moreover, our method significantly outperforms current research and commercial devices that do not use EEG (i.e., “EEG-less” methods). Finally, we validate the expert-level performance of a real-time scoring option.

The implications of these findings extend far beyond the technical achievements of the model itself. Our study shows the feasibility of using ECG for reliable sleep staging. We propose that this method will open new pathways for non-invasive sleep monitoring, with notable implications for patient care and sleep research.

2. Methods

2.1. Sleep datasets

In order to build a broadly applicable model, we used data from five large datasets from the National Sleep Research Resource [14]: The Cleveland Children’s Sleep and Health Study (CCSHS) [15] included 515 pediatric PSGs. The Cleveland Family Study (CFS) [16] included 730 PSGs from a wide age range. The Childhood Adenotonsillectomy Trial (CHAT) [17] included 1,639 pediatric PSGs. The Multi-Ethnic Study of Atherosclerosis (MESA) [18] included 2,056 PSGs from older subjects. The Wisconsin Sleep Cohort (WSC) [19] included 3,671 PSGs from middle-aged to older subjects. These datasets consist of PSG recordings scored using either R&K (CCSHS, CFS, and WSC) or AASM (CHAT and MESA). Finally, in addition to the wide range of ages (5 to 90 years), these datasets provide diversity in sex, race, ethnicity, and medical conditions.

As we will detail later, we processed each recording from the original studies to calculate various data quality measures (e.g., signal source, missing data, artifacts, etc.) and to harmonize the data (e.g., standardize the sampling rate and normalize the amplitudes). We discarded recordings that did not meet the quality metrics (defined later). However, we did not exclude any recordings based on subject characteristics (i.e., demographics, health, medications, etc.) or their sleep composition (i.e., time spent asleep or in any particular stage).

Next, our pipeline selected 4,000 recordings at random (3,000 for training, 500 for validation, and 500 for testing). We aimed to have the randomly selected distribution of subjects match the 2022 U.S. census estimates across decades as well as the mean age (Fig. 1a). Additionally, we mirrored the distribution of age, sex, and recording source across the three sets (i.e., training, validation, and testing). Given the variations in demographics of the source datasets, there are variations in the decade distributions for each study (Fig. 1b). As desired by selecting a broad range of subjects, the chosen recordings contained a wide distribution in stage ratios (i.e., the percent of the night spent in a specific stage), including a substantial number of recordings that had no epochs containing N1, N3, or both (Fig. 1c). Additionally, there is substantial variability in recording lengths, from 5.5 to 14.3 hours (Fig. 1d). When stratifying the recordings by decade, there are noticeable age-dependent shifts in the stage ratios (Fig. 1e). We expected the likelihood of time-dependent changes in stage ratios (e.g., N3 is more common early in the night, and REM more common at the end) and the age-dependent shifts (e.g., N3 decreases with age).

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

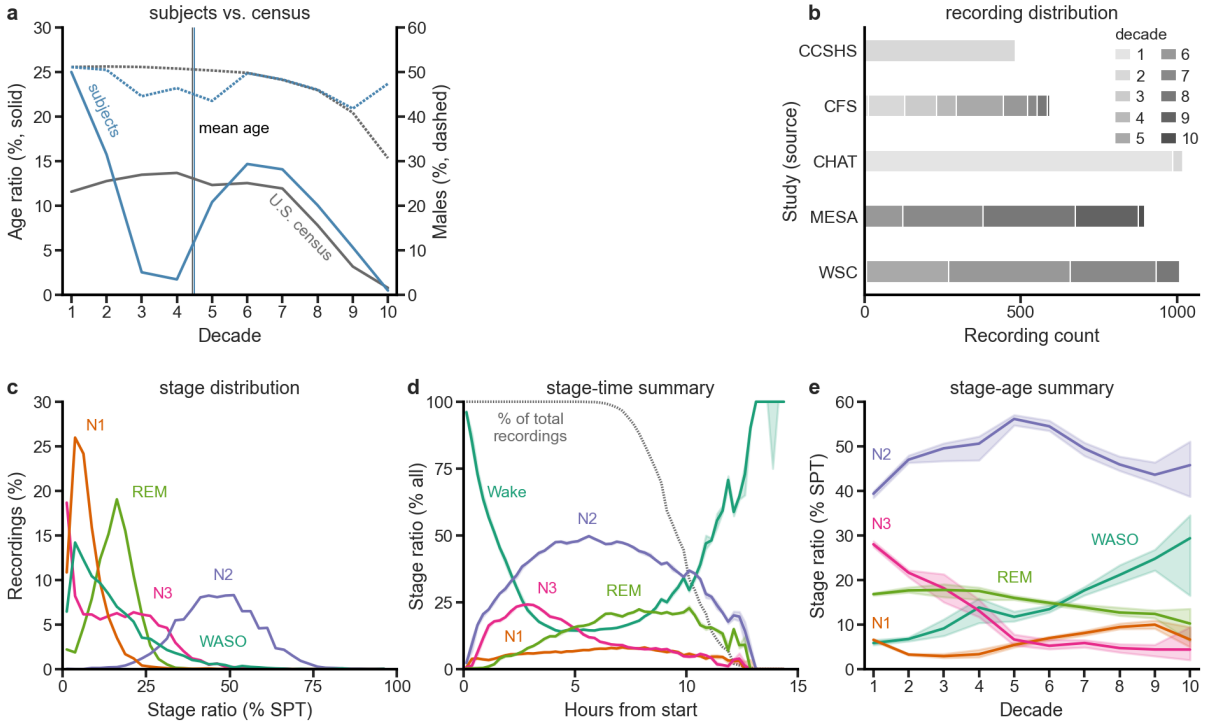


Fig. 1. Sleep datasets statistics

(a) We aimed to select subjects (blue lines) to match the U.S. census statistics (gray lines) by age (solid lines) and sex (dashed lines). The lack of subjects in decades 3-5 is a limitation of the available datasets (decade 1=age 0-9yr.), with subjects added to other decades to achieve the same mean age as the census data. (b) The 4,000 recordings came from five studies, with the distribution of the subjects' ages in decades shown. (c) There is a wide distribution of stage ratios as a percentage of sleep period time (SPT, i.e., the period between and including the first and last epoch of sleep). Wake after sleep onset (WASO) is any wake (arousals) during SPT. (d) As expected from previous studies, recordings show time-dependent changes in the relative proportion of the various sleep stages. E.g., N3 is more common at the beginning, and REM is more common at the end. e) The data also show expected age-dependent changes in sleep. In particular, the ratio of time spent in stage N3 declines with age, whereas arousals (WASO) increase.

2.2. ECG processing and selection

Even though each recording included data from many biophysical inputs, we used only ECG lead I (the limb lead across the heart) for our model's input. Some studies provided ECG lead I as a single channel of ECG data (often labeled "ECG" or "EKG"). Other studies separately provided recordings of the right (RA) and left (LA) limb electrodes. For those studies, we calculated ECG lead I by subtracting the two electrodes from each other (i.e., RA-LA).

2.2.1. Pre-processing

Because ECG is often an afterthought for PSG collection, there was a considerable variation in data quality in the ECG recordings (e.g., intermittent or poor connections, different sampling rates, and environmental noise). Therefore, we had to process and evaluate all 8,611 recordings provided by the five datasets to determine which recordings we could use to train and evaluate the model. To that end, all recordings went through an automated pre-processing algorithm described below. It bears mention that we took the recordings as-is and did not trim wake periods before the subject fell asleep (mean sleep latency = 1.3±1 hr.—one SD, see Supplementary Fig. S2f) or after

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

the subject woke up. Additionally, among the datasets, there were recordings in six different sampling rates (100, 128, 200, 250, 256, and 512 Hz). Therefore, we had to resample them to a common frequency (256 Hz).

1. If the ECG length was not a multiple of 30 seconds, we trimmed it down to the next nearest 30-second epoch length.
2. We silenced (i.e., set the signal to a value of 0) sections affected by intermittent connections. If we derived the ECG lead from two electrodes, we silenced both signals whenever there was a connection issue with either.
3. If we used two electrodes, we subtracted one from the other to obtain ECG lead I.
4. High-pass filter (0.5 Hz) the data to attenuate baseline wander but maintain longer features, such as T waves.
5. Remove 60 Hz line noise with a notch filter.
6. Remove any additional, automatically-detected, constant-frequency noises using notch filters.
7. Resample to a single common frequency (256 Hz).
8. Normalize using a robust z-score.

2.2.2. Detecting heartbeats

Once we had pre-processed every recording, the next step was identifying most heartbeats in each recording. The first pass involved finding heartbeats based on archetypical heartbeat templates. Next, we generated a recording-specific template for each recording from the first pass of detected heartbeats. Finally, we performed a second pass to add or remove heartbeats that matched the recording-specific template.

After identifying the putative heartbeats for each recording, we could calculate a recording-specific normalization factor, which brings all recordings to the same scale. To calculate the normalization factor, we first calculated the maximum absolute value for every identified heartbeat. Next, we took the 90% percentile of all maximum values as the maximum threshold. We set the normalization factor as twice the threshold to account for significant amplitude variations while maintaining a sufficient range. After that, we divided the ECG obtained after pre-processing using this normalization factor. The result will be that all, or most, heartbeats will fall within the range of ±0.5. Finally, to eliminate extreme values, anywhere the ECG exceeds ±1, we clipped the values to ±1. We clipped the values because neural networks work less well on data with extreme ranges.

2.2.3. Acceptable recording criteria

We wanted to ensure that all data used to train and evaluate the model were of decent quality (i.e., an allowance for reasonable quality variations without training the network on garbage data). Therefore, we only set selection criteria based on the ECG data. In other words, none of the criteria were based on the stage scores (e.g., time spent awake, etc.).

1. At least 5 hours of data, but no more than 15 hours.
2. Contain the lead I ECG channel or the two electrode channels necessary to derive it.
3. A sampling rate of at least 100 Hz.
4. There were three or fewer constant-frequency noises (including 60 Hz).
5. At least 85% of the epochs must contain some signal (defined as having more than eight unique values).
6. At least 85% of the epochs must contain at least one template-matching heartbeat.
7. At least 85% of the epochs must contain median absolute deviation (MAD) values ≥ -3 SD of the robust z-scored MAD values for all epochs.

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

8. The normalization factor must be ≤ 250.
9. After dividing by the normalization factor, ≤ 5% of the data can be extreme values (i.e., values outside ±1).

2.2.4. Recording selection and set building

At this point, there were 5,718 recordings remaining (of the original 8,611) which met the above selection criteria. Unfortunately, there were insufficient recordings from subjects in their 3rd, 4th, 5th, and 10th decades to match the U.S. census estimates. Therefore, we over-sample subjects from the remaining decades with the following two goals. First, the mean age should match that of the U.S. census. Second, the subjects from the 6th to 9th decades should be over-sampled by the same number. Even with these changes, we still desired to match the sex distributions for each decade, as provided by the census data.

To keep training times reasonable, we selected 4,000 recordings that we would use for the training, validation, and testing sets. We used random sampling to select the 4,000 recordings from the 5,718 available—with the age and sex distributions specified above. Because there are more recordings than unique subjects, we put additional criteria in the random selection process. First, the 500 recordings in the testing set must come from 500 unique subjects. Furthermore, we allowed more than one recording from the same subject for the training and validation sets on the condition that the subject was only in a single set. Finally, because random selection will probably draw from the original datasets unequally, the last step was to shuffle recordings between the sets to achieve similar dataset ratios. It bears stressing that we did not add the unselected 1,718 recordings to our testing set. To do so would have skewed the age and sex distributions away from the desired census distribution. We discuss these additional recordings later in Methods 2.14.

The 4,000 recordings consisted of 4,597,343 epochs (38,311 hours) of data (each recording duration = 9.6±1.4 hr.). See Supplementary Table S6 for the set-specific counts and Fig. 1 for visual representations of the recording statistics.

2.3. Sleep scores and weights

The source datasets provided the sleep score annotations in various file formats, using either R&K or AASM scoring criteria. Although slight differences exist in the similarly named stages, we harmonized the annotations with the following two adjustments to the annotated scores. When a dataset scored with R&K provided separate S3 and S4 stages, we combined them into a single stage: slow wave sleep (SWS/N3). Second, if the human scorer had annotated an epoch as anything except the five stages of interest (e.g., “unscored” or “movement”), we changed the epoch’s score to Wake and set the weight to zero. In total, 1.2% of the epochs fell into this bucket. By setting the weight to zero, we did not penalize the network in these cases, as there was no stage to compare with. In addition, because the human annotator had not scored these epochs as one of the five stages, we excluded them from all results.

We also adjusted the epoch-specific weights when the data quality of the epoch was poor. If we had silenced a portion of an epoch due to intermittent connection, we set the weight to one minus the proportion removed. Therefore, if we had silenced an entire epoch, it received a weight of zero. Finally, if an epoch contained eight or fewer unique values (i.e., distinct voltage readings), we considered it devoid of signal and assigned a weight of zero. In all, 0.9% of the epochs had no ECG data. In contrast to the unscored epochs above, we still included these epochs in the kappa calculations—even though the epochs were devoid of data.

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

2.4. Neural network

A neural network, mimicking the human brain's structure, recognizes patterns and solves complex problems by processing inputs through layers of interconnected nodes or neurons. The input for our neural network is the entire ECG recording, the subject's age and sex, and the recording time. The network outputs the probability for each sleep stage for all epochs at once (Fig. 2). Conceptually, we divided the network's layers into three groups. The first group, the feature extraction layers, extracts relevant features from the input data for each epoch. The second group, the temporal fusion layers, combines the features temporally across epochs. The third group, the classification layers, uses the fused features to assign a probability for each stage of each epoch. It is worth mentioning that the final structure presented here was arrived at through hundreds of hyperparameter search iterations. In other words, the decisions we made were through trial and error.

First, the feature extraction layers take the input to produce a set of features. The inputs include the cleaned, but otherwise untransformed, ECG (described above), the subject's sex (Boolean value) and age at the time of recording (normalized to 1 = 100 yr.), epoch's relative location within the recording (-1 = beginning, 1 = end), and wall time (where 0 = midnight and ±1 = ±24 hr.). The wall time is included because circadian rhythms influence stage proportions [20]. The ECG data passes through several convolutional and pooling layers to extract 40 features. The network combines those ECG features with the subject's age and sex, as well as the time variables. Finally, several dense linear layers reduce the output to 25 features for each epoch.

Next, the temporal fusion layers take the 25 features for each epoch and merge them across time. The configuration is based on the temporal convolution network [21], with a modification to merge information before and after the current epoch. The first layer merges the features from the current epoch and epochs before and after it (epoch±1). The second layer merges the already-merged features from the current epoch and the epochs located two positions before and after it (epoch±2). This motif continues upward for 12 layers as epoch$\pm 2^{(\text{layer}-1)}$, such that any epoch in the top layer could combine information from epoch±34 hr. This width is substantially longer than any recording used or ever expected.

Finally, the classification layers take the 25 features from the final temporal fusion layer through several layers of dense linear units. The output of the final softmax layer resembles the confidence or probability for each of the five sleep stages. We take the epoch's sleep stage as the stage with the highest probability.

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

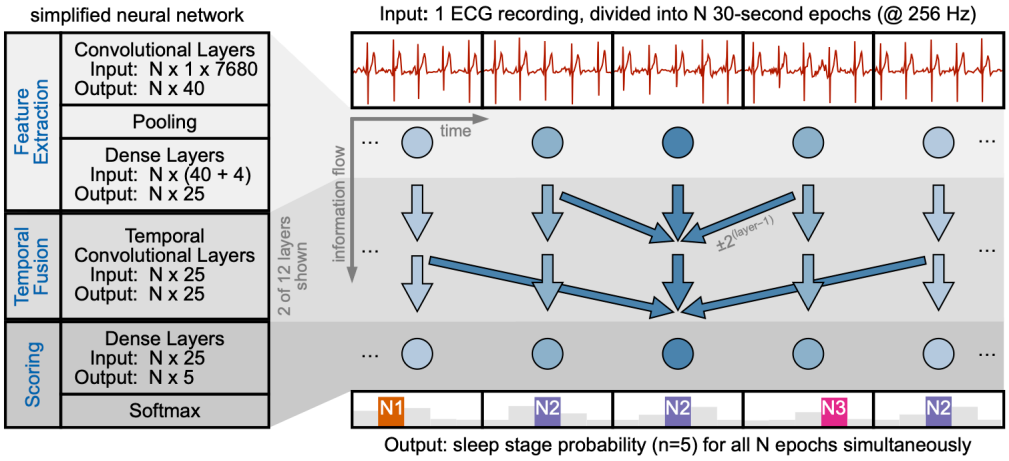


Fig. 2. Neural network

The network consists of three groups of layers. It takes a single recording of ECG data as input and scores all epochs (N) at once. Given the variable length of recordings, ellipses represent a duplication of network structure across the recording. The four other inputs (not shown) are age, sex, wall time, and relative epoch position within the recording. We hid most arrows for clarity, and ECG and epoch stage scores are for illustrative purposes only.

2.5. Training and evaluation

Designing a neural network involves a hyperparameter tuning process (e.g., determining the optimal number of layers, node types, non-linearities, etc.). During the hyperparameter search, we trained the model on only the training set (n = 3,000) and only evaluated it on the validation set (n = 500). We used a separate hold-out testing set because cross-validation with that data (i.e., partitioning the data such that a network eventually evaluates every sample) is inappropriate during hyperparameter turning. A more thorough discussion on the information leak that happens when using cross-validation is in Supplementary Discussion 6.3.2. A more detailed description of the history of our hyperparameter search is in Supplementary Methods 6.1.1.

In addition to standard techniques to improve regularization, we used automatic per-recording weighting during training. The goal was to lower the weight of recordings with significantly lower kappas. We weighted the recordings to prevent the network from misusing parameters towards improving classification on just a handful of recording outliers. Additionally, we identified from the waveforms that the technicians occasionally attached the ECG leads to the wrong terminals. Therefore, to improve the model's robustness to inverted leads, we had the data loader invert the ECG of every recording with a 50% probability during training. We did have the loader invert the signal during the evaluation phase.

The network was trained using a variant of the Adam optimizer with stable weight decay, AdamS [22]. The training algorithm reduced the learning rate by half when performance had plateaued for 50 training epochs (i.e., complete runs over the entire training set) on the evaluation set. The initial learning rate was $1x10^{-3}$, with a minimum of $1x10^{-6}$. We trained the model using a batch size of 10 (i.e., the number of training recordings grouped together) for a maximum of 1,000 training epochs. We stopped the hyperparameter search and selected the final model when we decided that additional changes were unlikely to materially improve the results on the validation set.

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

After we selected the final model, we retrained it on the joint training and validation set (n = 3,500). At this point, we evaluated it on the hold-out testing set (n = 500). We had never used the hold-out testing set before.

2.6. Performance metric: Cohen's kappa

To compare our model's performance on scoring sleep stages, we need to use an established metric. When classifying sleep stages, there is some subjectivity and no "ground truth" score for any given epoch. In other words, even identically trained sleep scorers will have some disagreement [6], [7]. Therefore, the most appropriate statistical measure is the inter-rater agreement, or the degree of agreement between two observers. We used Cohen's kappa (κ), the most commonly reported measure of inter-rater agreement in sleep research.

Cohen's kappa measures the agreement between two raters that one cannot attribute to chance alone. The statistic, shown below, uses the probability of observed agreement (p_o) and the probability of chance agreement (p_e). A value of $\kappa = 0$ means no agreement above chance level, while $\kappa = 1$ means perfect agreement—where both raters match on all annotations. The minimum value depends on the marginal distributions and is between zero and -1. Additionally, when the relative prevalence of one or more classes is sufficiently far from balanced, the overall and stage-wise kappas will decrease. This decrease is a natural consequence of the measure [23].

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

While designed for binary classification, it is possible to compute the individual class-specific kappas in a multiclass task such as sleep staging. For a multiclass task, the contingency table is split into separate class-versus-others tables for each class [24]. For instance, to compute kappa for N3, the first class is N3, and the second class is the combination of the other classes (Wake+N1+N2+REM). Unfortunately, if there are two or more classes, and both raters score everything as just one of the classes, the naïve formulation will produce an undefined value. However, in this case, both raters perfectly agree (i.e., $\kappa = 1$); therefore, we set kappa to 1.

Depending on the field, two different methods of computing and presenting kappa exist. Therefore, we must do the same to compare our results with others. The first method, often used by human-scored PSG literature, computes kappa for each recording individually and calculates summary statistics on those kappa values. The second method, favored by machine learning literature, aggregates one contingency table for all epochs from all recordings and computes kappa on that aggregated contingency table. We found that when the number of epochs and recordings are both large, the median kappa of all recordings is similar to the kappa of all epochs. However, for all results, we specify which method we used.

2.7. A new kappa-correlated loss function

A loss (or objective) function is a mathematical function that quantifies the difference between the predicted outputs of a model and the actual target values. Its purpose is to guide the training of the neural network by minimizing this difference, i.e., to calculate the gradients used to adjust the network's weights and biases. Our criterion for the loss function was the highest overall kappa with the narrowest possible range of individual stage-specific kappas. In other words, every stage-specific kappa should be as high as possible, instead of the typical outcome where the classifier ignores the minority classes.

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

For a classification task such as sleep staging, cross-entropy loss (aka log-likelihood) is the standard [25]. However, the first issue we had with cross-entropy was that it assumed that the classes were nearly equal in size. When class size imbalances naturally exist, such as with sleep stages, cross-entropy often disregards the minority classes. The issue is glaring for N1, which often constitutes less than 5% of the night but is still an essential marker of the wake-sleep transition. Various techniques are employed to overcome this issue, including under- and over-sampling the classes [26]. However, a data-level solution was not possible here because our model scores an entire night of sleep as one “unit”. The constraint of operating on an entire recording at once makes it impossible to balance the proportions of the classes artificially. Another solution often used is weighting each class's importance. Extensive evaluations by us found that weighting was also inadequate; although the overall kappa slightly increased, it had a negligible effect on the kappa of N1.

The second related issue with cross-entropy loss is that it assumes accuracy is the suitable performance metric. However, accuracy is only loosely correlated with kappa. Although $\kappa = 1$ when accuracy is 100%, there is otherwise no fixed one-to-one mapping. Moreover, the kappa value will depend on the relative proportions of the classes. For instance, when accuracy is 50%, kappa could be almost any value between -1 and 1. That is to say, for the same accuracy, kappa could be wildly different.

Due to the limitations of existing loss functions, we developed a new function. Our loss function is one minus the geometric mean of the scaled class-specific kappas. Therefore, we call it the class kappa mean. As expected, our loss function is (negatively) correlated with the overall kappa, which is the weighted arithmetic mean of the class-specific kappas. An advantage to using the geometric mean is that it is more invariant to the relative proportion of the classes, i.e., the function is less likely to ignore the minority class. Since kappa can technically range from -1 to 1, the formula re-scales the kappas to stay within the range [0, 1]. This scaling prevents issues that could arise with a kappa less than or equal to zero. For the formula below, *c* is the current class, and *n* is the number of classes (for five-stage scoring, $n = 5$).

$$\mathcal{L} = 1 - \left(\prod_{c=1}^{n} \left(\frac{\kappa_c + 1}{2}\right)\right)^{\frac{1}{n}}$$

We calculated the kappas for the loss function using a contingency table generated from the output of the softmax layer (described above in Neural network, 2.4). During training, the algorithm builds a contingency table, which enables the back-propagation algorithm (using the loss function) to improve the network. The loss function causes the network to be more confident in the epoch scores that agree with the human scorer. Additionally, it minimizes the likelihood and confidence of epoch scores that disagree with the human scorer. Our loss function was critical for training the neural network.

2.8. Meta-analysis of human-scored PSG

We used a recently published meta-analysis on inter-rater agreement of sleep staging as a starting point [7]. In examining the input data used in this meta-analysis, we noted discrepancies between the values reported in the original studies and those utilized in the published meta-analysis.

First, a handful of studies were included that had only a single kappa value or contingency table (i.e., no variance was provided, i.e., SE or SD) [27], [28], [29], [30]. While there is a formula for

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

estimating the variance (SE) for kappa based on a single contingency table [31], this value is conceptually different. It is the variance expected by chance of the two scorers' scores, not the variance of kappas from multiple scorers. We need the variance of kappas from multiple scorers for a meta-analysis on inter-rater agreement. Therefore, we removed those studies with only a single kappa or contingency table. Additionally, some of the studies only provided boxplots of their results. Using published techniques, we converted these quartiles into mean [32] and SD [33] for our meta-analysis. We tabulated all these details, their source within the papers, and notes for the overall kappa in Supplementary Table S1, and for each stage-wise kappa in Supplementary Table S2.

Our meta-analysis used the DerSimonian-Laird random-effects model [34] to obtain the kappa estimates for each stage and their 95% confidence intervals (CIs). We calculated I^2 to assess heterogeneity. Although the CI estimates the kappa range for the pooled result, the 95% prediction interval (PI) gives the likely range of expected future studies of the same type and is usually wider than the 95% CI. We used a bootstrapping method to calculate the 95% PIs [35].

Finally, we created funnel plots and numerically tested the included studies for publication bias [36].

2.9. Comparison with human-scored PSG: non-inferiority testing

Our claim is that our model's performance is on par with the current standard, human-scored PSG. To substantiate this claim, we employ non-inferiority testing. Researchers and clinicians use this testing approach to demonstrate that a new method or treatment is not inferior to an established standard by more than a clinically relevant margin. This method contrasts with the common null hypothesis significance testing (NHST), which seeks to identify significant differences without specifying a minimum performance level. Non-inferiority (NI) testing is extensively used in fields requiring rigorous validation of new interventions or models [37], such as clinical drug trials and, more recently, machine learning in sleep research [38]. This method ensures that innovations are at least as effective as current standards without compromising performance or utility.

For our study, the null hypothesis posits that our model's performance (i.e., Cohen's kappa, μ_{model}) is below the inter-rater agreement of human experts using PSG ($\mu_{standard}$) by a clinically relevant margin (Δ_{NI}). Conversely, the alternative hypothesis suggests that our model's performance is not significantly worse, demonstrating non-inferiority.

$$H_0: \mu_{model} \leq (\mu_{standard} - \Delta_{NI})$$
$$H_1: \mu_{model} > (\mu_{standard} - \Delta_{NI})$$

Sometimes, statisticians explain the same in terms of a performance threshold instead. This performance threshold is equal to the mean of the standard minus the non-inferiority margin, or $threshold_{NI} = (\mu_{standard} - \Delta_{NI})$.

$$H_0: \mu_{model} \leq threshold_{NI}$$
$$H_1: \mu_{model} > threshold_{NI}$$

Specifying a threshold the new method must meet or exceed is critical and is determined based on clinical judgment and statistical considerations. Researchers usually set it as a percentage of the standard's lower 95% CI. The higher the percentage, the more rigorous or conservative the threshold for non-inferiority. Typical thresholds range from the most common, 50%, up to 90% for the most rigorous and critical evaluations, such as those with mortality considerations, e.g., new

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

antibiotics. This non-inferiority margin reflects the smallest difference in performance that is relevant in practice, ensuring that the new method's efficacy (as measured by its 95% CI) is meaningfully comparable to the established standard (as measured by its 95% CI).

We adopt the most rigorous threshold possible here; namely, we set our threshold at 100% of the lower 95% CI of the meta-analysis. This threshold ensures that no portion of the 95% CI of our model's performance (i.e., kappa) falls below the 95% CI of the random-effects estimate.

The alpha level for a non-inferiority test when using the 95% CIs is 0.025. This lower-than-expected value is standard [37]. This stringent alpha level ensures a rigorous assessment of non-inferiority.

Because we will perform multiple tests with each stage's meta-analysis estimate, we use the preferred Hochberg step-up procedure [39] to control for the family-wise error rate. In other words, we adjust the p-values for multiple comparisons, and the adjusted p-values further decrease the accidental (i.e., chance) finding of significant (i.e., non-inferior) results.

2.10. Comparison with EEG-less models

To compare our results with published non-PSG EEG-less results, we had to review the literature to determine which studies we could include. While the current literature on EEG-less methods (including current commercial sleep-tracking devices) is more extensive than what we included, we excluded papers for one or more reasons. The possible reasons include not listing kappa, having a kappa below 0.5, only considering two-stage scoring (i.e., Wake/Sleep), or their evaluation set was not independent of their training set. In other words, for the last point, the study must have a hold-out testing set to act as an unbiased estimate for future unseen data. Please see Supplementary Discussion 6.3.2 for additional details on the methodological issues we noticed. We list the studies we compared with in Supplementary Table S5.

Most EEG-less studies compute a single kappa value on a single aggregate contingency table of all epochs for their entire testing set (i.e., ignoring recording-by-recording differences). Therefore, we calculated kappa on a bootstrapped aggregate contingency table to compare our five-stage model with the other EEG-less models. We performed the bootstrap by sampling with replacement a sample of 500 recordings from the testing set 10,001 times. For each bootstrap sample, we computed a single contingency table of epochs from the recordings in the sample. Using percentile bootstrap, if the next-best model performs below the lowest bootstrapped result, we can only conclude that the p-value is below the inverse of half the bootstrap iteration count (i.e., 1/5,000).

However, since most models reported in the literature do not evaluate five-stage scoring, we also aggregated the stages from the five-stage model's results. For example, for four-stage scoring, we set the epochs of either N1 or N2 to "Light" sleep. We also conducted the same bootstrap operation described above to compare the performance of four- and three-stage scoring.

2.11. Sleep scoring concordance

To compare the human-score stages with the model-scored stages, we used a row-normalized contingency table to show where there is the most substantial agreement and disagreement. We plotted an embedding to look at the feature clustering that the network builds to score sleep. Specifically, we used a t-distributed stochastic neighbor embedding (t-SNE) to reduce the 25 features to two—which we can more easily plot. Doing so allowed us to visualize clusters in the

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

embedding space. Furthermore, we can visualize disagreements in this 2-D representation by highlighting the epochs in which the human and the model disagree. In addition to the overall kappa (i.e., agreement) between the human and model, we also investigated the transitions between stages—specifically, the transition rates and probabilities. Moreover, we compared the human and model transition matrices by bootstrapping and using Pearson's r to calculate a correlation between them.

2.12. Robustness to noise and other perturbations

Given the limitation that we could not generate synthetic data, we had to conduct other experiments to analyze the robustness of the network. These experiments included trimming recordings, adding noise, modifying the recording's epoch order, silencing whole (and portions of) epochs, and changing the demographics. We trimmed the recordings by removing epochs from the beginning or end. For the noise experiments, we individually added four different sources of noise: white Gaussian noise, as well as three noises from the MIT-BIH noise stress test [40] (i.e., baseline wander, electrode movement, and movement artifact).

Several of these experiments allow us to investigate how much contextual information matters. Silencing (or setting to zero) whole epochs was the first experiment. We randomly selected epochs in each recording and silenced their ECG input. Another experiment to investigate the contextual information was either flipping the epoch order of each recording or shuffling the order. Each epoch would still have the ECG progressing through time in the expected direction, but adjacent epochs would now be out of order. Moreover, using the method of integrated gradients [41], we can investigate how important each epoch in a recording is for scoring a given epoch.

Finally, we evaluated the remaining two inputs: sex and age. By either flipping the sex or assigning a random age, we can determine how important those variables are to the performance of the network. Related to the integrated gradient analysis of the epochs, silencing portions of all epochs allowed us to investigate what portion of the epoch the network was using. Either we silenced the epoch from both ends or the center.

2.13. Additional model variations

As mentioned, most of the results presented come from just the single, primary five-stage model described above. However, we trained an additional eleven models. We used one model to evaluate a real-time variant. We used another model to determine a better approximation of a naïve classifier's performance. Finally, we used nine models to compare and assess the loss function we present here. We stress that the single, primary five-stage model above stands alone, and the additional models do not form an ensemble.

2.13.1. Real-time model variant

A scorer typically scores sleep after the night or recording is over. Therefore, the scorer has access to the entire night to assist them in classifying each epoch. However, real-time (i.e., causal) scoring would be necessary for some of the interventions and applications we think would benefit from cardiosomnography. To that end, we trained one five-stage real-time model. We created this model by modifying the "Temporal Fusion" section to remove access to information following the current epoch (Fig. 3). In other words, for any given epoch, it could not use information from the epochs that followed it (i.e., the future). This modification included removing (by replacing it with random data) the relative epoch position input, as this information reveals the remaining recording time.

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

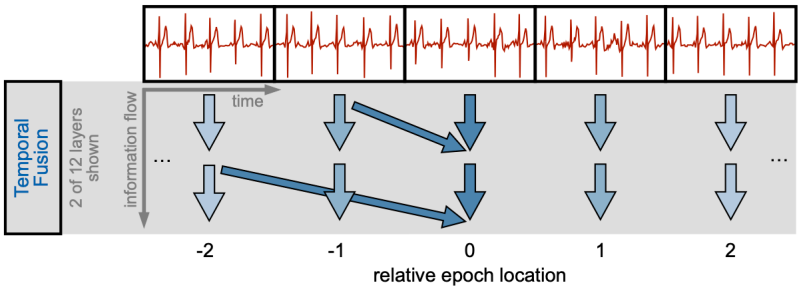


Fig. 3. Network modification for real-time scoring

For the real-time variant, we modified the "Temporal Fusion" section of the network (Fig. 2) to force the network to score in real-time (i.e., causally). Note how no arrows point anti-causally. We hid most arrows for clarity, and the rest of the network remains unchanged.

2.13.2. Determine a time-only floor

If one were to remove the ECG input entirely, it would be reasonable to suspect that the model would perform no better than chance ($\kappa = 0$). However, the fact that sleep stages occur in cycles and have an expected progression across the night raises the possibility that using only time might achieve better-than-chance agreement. We were unaware of literature evaluating this idea. Therefore, we decided to determine the performance using the time variables as the only input (i.e., wall time and relative epoch position). So that we could use the same network structure, we replaced the ECG, age, and sex input data with randomly generated data that changed each time. We trained (one) time-only five-stage model.

2.13.3. New loss function comparison

Although we developed our loss function during the hyperparameter search, it would be helpful to know how well other common loss functions perform on the same five-stage model. To that end, we trained the same five-stage model four times with commonly used loss functions. Furthermore, because of the tradeoffs that all loss functions exhibit, we wanted to determine what would happen if the same five-stage model only had to score a single stage (e.g., only Wake vs. sleep). Therefore, we also trained five individual one-stage models with our loss function.

2.14. Evaluation of additional hold-out recordings

As mentioned, there were an additional 1,718 recordings from the original five studies that met the acceptable recording criteria (Methods 2.2.3) that we did not use. We did this out of a desire to limit training time and to match the U.S. census distribution. However, we did evaluate these recordings separately to determine if there was anything unique about the recordings we had randomly selected.

Furthermore, we evaluated recordings from a dataset we did not use in the training phase, namely the MrOS Sleep Study (MROS) [42]. We analyzed these recordings to determine if the model was learning and using any study- or site-specific features from the five primary studies. The study provided 3,933 recordings, of which 3,193 met the quality criteria.

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

2.15. Miscellaneous common procedures

Where possible, we used nonparametric bootstrapping—a method that makes no assumptions about normality. For all bootstrapped results, we always used a sample of the same size as the original (e.g., n=500 for the testing set) with 10,001 iterations. Moreover, where possible, we use the median, a more robust measure of central tendency, as the estimator. Whenever line charts show shaded regions, the line is the median, and the shaded regions are the 95% CIs. We estimated the CIs for those figures using percentile bootstrapping of the median.

3. Results

We have organized our results into six sections. Initially, we compare our model's performance with expert human-scored PSG using the meta-analysis estimates. Additionally, we compare it with other EEG-less models (i.e., models that exclude EEG as an input) and illustrate its concordance with human-scored PSG. Further, we analyze the model's robustness to noises and perturbations. Next, we showcase its real-time capabilities. Finally, we cover the remaining miscellaneous results. These findings underpin the suitability of ECG for the highest-quality sleep studies and demonstrate a network that could make cardiosomnography widely available.

3.1. Comparison with human-scored PSG

To assess the claim that the model achieved expert-level human performance, we must compare its performance with how well human scorers agree with each other. First, we conducted a meta-analysis on human inter-rater agreement. Then, we performed non-inferiority testing against those estimates to assess the on-par claim.

To begin with, we conducted a series of meta-analyses using the data from eleven studies that assessed inter-rater agreement (i.e., between two human scorers) on human-scored PSG. These meta-analyses assessed the overall inter-rater agreement between human scorers (top of Fig. 4, source data in Supplementary Table S1) as well as the kappa for each stage individually (top of Fig. 5, source data in Supplementary Table S2). The results of these meta-analyses show that the overall inter-rater agreement is high and that there is some variation in the stage-wise agreement. We will discuss these stage-wise variations several times later. We have tabulated the complete meta-analysis results in Supplementary Table S3. In addition to the estimated kappa, we show the 95% CI (the expected variation in this estimate) and the 95% PI (the range of likely future studies). We also tested for publication bias using visual and numerical techniques and found no signs of bias in the meta-analysis inputs (see funnel plots; Supplementary Fig. S1).

It is worth succinctly repeating that although we list studies used in the meta-analysis chronologically, they do not, and indeed cannot, represent an evolution of human inter-rater agreement. That is to say, there is no expectation that the kappa values should increase with time. Consequently, no threshold is ratcheting up for what constitutes an "acceptable" human-scored PSG agreement. By definition, the inter-rater agreement between two expert human scorers is just the result of two scorers scoring the same recording. Instead, the natural and expected variation in inter-rater agreement is due to numerous factors, the foremost being the somewhat subjective nature of sleep stage scoring. The factors also include different sample sizes, patient populations,

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

scorer training, and equipment. It is this variation that required a meta-analysis in order to determine the range of the "average" human-scored PSG agreement.

Next, for each stage, we tested our results for non-inferiority (i.e., on par or better than performance). We present the results of our primary model on the entire testing set below each of the meta-analysis results (Fig. 4 and Fig. 5). We also show the results of the non-inferiority tests in the same area, with significance bars marking every significant result (i.e., non-inferior). When we evaluate our primary model against the meta-analysis results, all comparisons are non-inferior except for N3. We used the Hochberg procedure to correct for the multiple comparisons with the random-effects estimate for each stage. The alpha level for a non-inferiority test using the 95% CIs is 0.025. The complete tabulation of the adjusted p-values is in Supplementary Table S4.

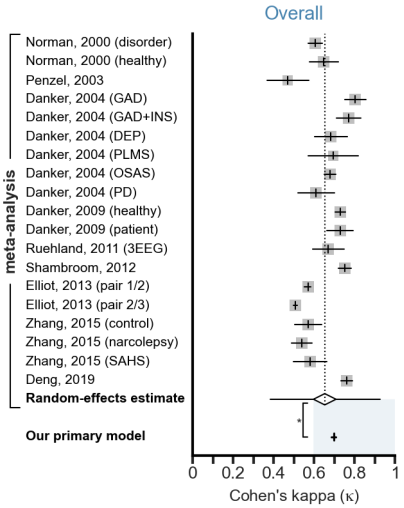


Fig. 4. Forest plot for overall kappa and our model's performance

Above is the forest plot for the random-effects estimate of the five-stage Cohen's kappa (κ) for human-to-human inter-rater agreement on PSG. We list the source inputs chronologically, with each mean kappa and CI to the right. Per convention, the gray square for each represents their weight. The unfilled black diamond represents the random-effects estimate (width representing the estimate's 95% CI), and the whiskers extending from the diamond represent the 95% PI. The black vertical dotted line also indicates the random-effects estimate. Below is the five-stage kappa of our primary model based on single-lead ECG evaluated on our testing set. The non-inferior comparison using a t-test is significant (i.e., non-inferior; *$p<0.025$). We tabulated the study source data in Supplementary Table S1, the meta-analysis results in Supplementary Table S3, and the adjusted p-values in Supplementary Table S4. We found no publication bias (Supplementary Fig. S1).

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

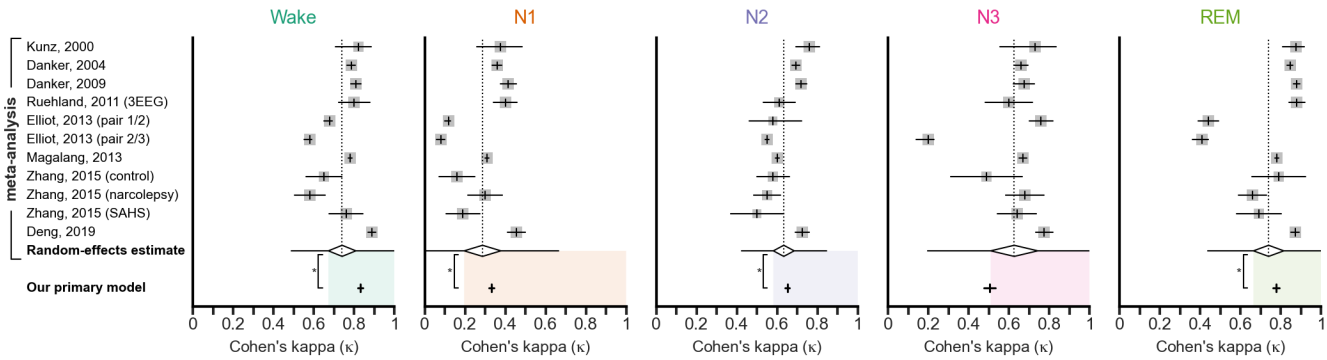


Fig. 5. Forest plots for stage-specific kappas and our model's performance

Above are the forest plots for the random-effects estimates of each stage-specific Cohen's kappa (κ) for human-to-human inter-rater agreement on PSG. We list the source inputs chronologically, with each mean kappa and CI to the right. Per convention, the gray square for each represents their weight. The unfilled black diamonds represent the random-effects estimates (width representing each estimate's 95% CI), and the whiskers extending from the diamonds represent the 95% PIs. The black vertical dotted lines also indicate the random-effects estimates. Below are the five-stage kappas of our primary model based on single-lead ECG evaluated on our testing set. All non-inferior comparisons using a t-test are significant (i.e., non-inferior; *p<0.025)—except for N3 (not non-inferior). We tabulated the study source data in Supplementary Table S2, the meta-analysis results in Supplementary Table S3, and the adjusted p-values in Supplementary Table S4. We found no publication bias (Supplementary Fig. S1).

As stated above, for overall and each stage-wise kappa, the only non-significant result (i.e., not non-inferior; $p > 0.025$) was for N3. To investigate this further, we performed the same comparisons, but now disaggregated by dataset (Fig. 6). We then found that, for overall kappa, the only non-significant result (i.e., $p > 0.025$) was for our primary model evaluated on WSC only. However, there were three non-significant results for the stage-wise kappas—all for stage N3. Specifically, the non-significant results for N3 include our primary model evaluated on the entire testing set and specifically on MESA and WSC only. However, the N3 performance was non-inferior when evaluated on the other three datasets. We used the Hochberg procedure to correct for the multiple comparisons with the random-effects estimate for each stage.

Investigating the N3 finding further, we found that the human-scored N3 stage ratios are markedly lower for MESA and WSC versus expectations for the same ages (Supplementary Fig. S3a). We also found the same trend for the model-scored results (Supplementary Fig. S3b), indicating agreement between the two scorers on the N3 ratios for those datasets. The dataset and decades with low N3 ratios also had lower N3 kappas (Supplementary Fig. S3c). Since, as mentioned, when the proportions of individual classes approach all or nothing, kappa will tend towards zero [23]. This finding is similar to the results when we stratify the stage-wise performance by decade (Supplementary Fig. S2a). The model's performance is largely unaffected by age except for N3, which decreases beginning in the fifth decade. Almost all of the older subjects come from the MESA and WSC datasets (Fig. 1b). This age-dependent result is expected and reported elsewhere [43], likely because, with age, sleep becomes shorter and more fragmented [44].

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

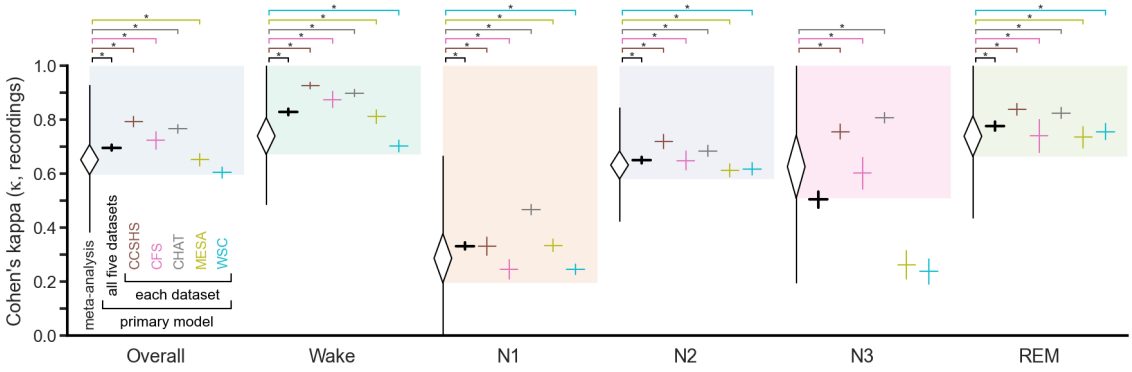


Fig. 6. Performance disaggregated by source dataset

For overall and each stage, we duplicated the meta-analysis estimate and our primary model's result on all five datasets to aid the reader (from Fig. 4 and Fig. 5). Then, for each stage, we disaggregated results by source dataset. The colored area is the non-inferior region. Almost all non-inferior comparisons using a t-test are significant (i.e., non-inferior; *p<0.025). The exceptions are Overall when evaluated on WSC only (not non-inferior) and N3 when evaluated on either MESA or WSC (both inferior). We tabulated the meta-analysis results in Supplementary Table S3 and the adjusted p-values in Supplementary Table S4.

3.2. Comparison with other EEG-less models

To assess the claim that our model achieves better performance than other EEG-less models and devices, we must compare our results with published non-PSG EEG-less results. Unfortunately, since there is enormous variability in the quality of and inputs used for these studies, we discuss the exclusion criteria in Methods 2.10 and Supplementary Discussion 6.3.2.

When comparing our model with other recently published EEG-less models and devices—including the state-of-the-art—we find it performs significantly better (Fig. 7, percentile bootstrap $p < 0.0002$). This finding includes when comparing against the best-published five-stage scoring ($\kappa = 0.726$ versus 0.585 [45]—evaluated on all epochs). Because most published models use more coarse stage granularity, we also evaluated the final model by combining the appropriate stages (Methods 2.10). Furthermore, our model performs better regardless of the granularity of stages (e.g., five-, four-, or three-stage scoring) or the number of additional inputs used (e.g., actigraphy, respiration, HRV, etc.; Supplementary Table S5).

To put the magnitude differences into context, we also include a "time-only floor", which represents our same model structure that is blind to the ECG, age, and sex. This floor approximates what one could expect from a naïve classifier that is only aware of the epoch's position in the recording—and is significantly above the expected $\kappa = 0$ threshold (i.e., chance agreement knowing only the prior probabilities).

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

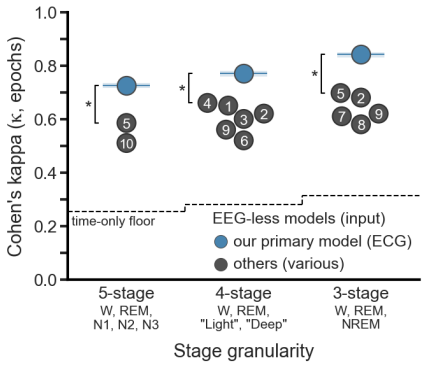


Fig. 7. Comparison with EEG-less models

Our model (blue dots and horizontal lines) performs significantly better than other EEG-less models (black dots with numbers). This result was true regardless of stage granularity—even while using fewer inputs (percentile bootstrap *p<0.0002). The bootstrapped 95% CIs were smaller than the dot diameters (shown as shading). The other dots represent the kappas of the recent best (κ≥0.5) EEG-less models. The sources are: 1) Radha [46] 2) Wulterkens [47] 3) Fonseca [48] 4) Sridhar [49] 5) Sun [45], 6) Beattie [50], 7) Yoon [51], 8) Domingues [52], 9) Willemen [53], 10) Sady [54]. Note that our model uses ECG as the only input, while many other models also use respiration, actigraphy, or both (i.e., some models used even more biophysical data). A “time-only floor” indicates the performance of the current model structure when only using the time variables (wall clock time and relative epoch position), i.e., no ECG or demographic data. We tabulated all details in Supplementary Table S5.

3.3. Sleep scoring concordance

In addition to overall and stage-specific performance, we investigated other aspects in which the model’s outputs and internal representations (i.e., feature spaces) are in harmony (or incongruence) with expectations of human-scored PSG.

The first investigation looks at agreements and disagreements (e.g., when there is disagreement about the stage of a particular epoch, what stage the two scorers say the epoch is). Another valuable perspective for assessing concordance with expected results—not well captured in single kappa value for all epochs—is to look at the transition rates (e.g., how often does a particular stage transition to another stage in each recording.

The row-normalized contingency table (visually similar to a confusion matrix, but without the assumption of either scorer being correct) showed high agreement (Fig. 8a). The percent agreement between the human and model scores was 80.0%. The main disagreement was with N1, which the model sometimes scored as one of the stages that naturally precede or follow it in time, namely Wake and N2. This finding matches what has been reported elsewhere for human-scored PSG data [55]. Furthermore, we found this same expected pattern for every stage; when there was disagreement, the other scorer scored it as the stage that typically preceded or followed it.

We confirmed these findings by investigating the network’s feature space, specifically by creating a t-SNE plot (Fig. 8b) of the 25 features for each epoch that came from the penultimate layer of the network (i.e., before the final classification occurred). The t-SNE calculations reduce the feature space to two dimensions and place similar epochs in the 25-D space closer together in the 2-D space. We plotted the epochs where the human and model agree as dots, colored by stage. Compared with the agreements, the disagreements are less frequent (in n = 453,866

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

epochs, the model agreed with humans, while in n = 113,275 epochs, it did not). Therefore, we used a kernel density estimate to show the density of the disagreements. Like the row-normalized contingency table (Fig. 8a), the greatest disagreement occurs when scoring N1 epochs, especially near the boundary between N1 and N2. The disagreements also occur along boundaries between adjacent stages in a typical recording.

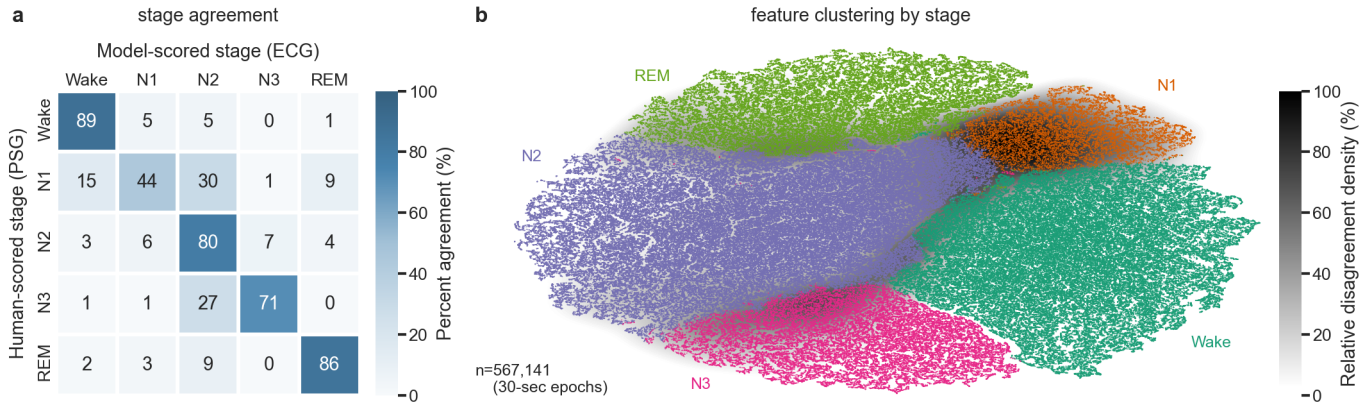


Fig. 8. Normalized contingency table and t-SNE of all epochs

(a) The row-normalized contingency table (of n=567,141 epochs) highlights the high degree of agreement between human- and model-scored stages (main diagonal). When disagreement exists, the model usually scores the epoch as the stage that typically precedes or follows it during the night, e.g., scoring N1 as Wake or N2, scoring N3 as N2, and scoring REM as N2. The lowest agreement occurs with N1, which is consistent with human scorers of PSG data. The percent agreement is 80.0%. (b) We generated the t-SNE plot from the 25 outputs of the penultimate layer of the network. Dots represent the epochs where the model agrees with the human-scored stage. Due to the sparsity of the disagreements, we used a kernel density estimate to show where the disagreements occur (black = highest disagreement density). The greatest disagreement density occurs when scoring N1 epochs. Disagreement typically exists along each of the cluster borders except Wake-N3.

In addition to assessing the model's performance (i.e., kappa), we also examined sleep stage transitions to assess the model's concordance with human scoring. The overall, Wake, and N1 transition rates are similar for the model and human (Fig. 9a). The overall transition rate also aligns with results reported by others [56]. Although the boxplots for the various stages largely overlap, N3 transitions show the greatest difference. Furthermore, the transition matrices for the model and human scores are nearly identical (Fig. 9b,c), with a bootstrapped Pearson's r of 0.998.

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

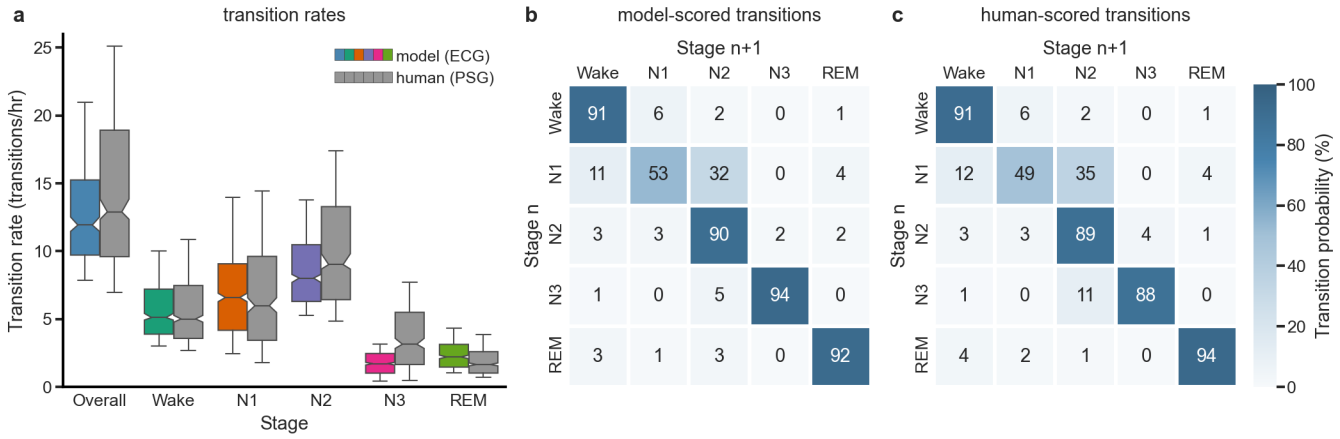


Fig. 9. Stage transitions

(a) Transition rates for the n=500 testing set for our ECG-based model (assorted colors) and the human-scored PSG (gray). There are minor differences in the Overall, Wake, and N1 transition rates. While there are significantly fewer transitions for the model for stages N2 and N3 and significantly more transitions for stage REM. (b) The transition matrix of the model's classifications for all n=567,141 epochs. (c) The transition matrix of the human's classifications (same n). The Pearson's r of the two matrices (panels b and c) is 0.998, with percentile bootstrapped 95% CI [0.997, 0.999]. In general, the next epoch is likely the same stage as the current one. Consistent with the literature, the probability that N1 will transition into a different stage (namely N2 or Wake) is higher than for other stages. Whiskers at P10 and P90.

3.4. Robustness to noise and other perturbations

Next, we investigated the robustness of the network and its suitability for more challenging environments by modifying the testing set and examining how those modifications affected performance. We modified the existing testing data because we cannot generate synthetic ECG data corresponding to a particular sleep stage. We stress that we conducted all these experiments with the final primary model and never trained the network on these manipulations (e.g., we did not add noise during training).

First, we examined the effect of trimming the recordings, either from the beginning (Fig. 10a) or from the end (Fig. 10d). When trimming from the beginning, the performance for every stage except N1 immediately trends downward, albeit to varying extents. When trimming from the end, however, the performance was nearly unaffected until after we removed many hours of data. The result underscores the greater importance of the initial period of the recording for classifying sleep stages (similar to the pre-sleep wake effects, Supplementary Fig. S2e).

To assess the model's ability to handle noise (e.g., environmental), we separately added four different sources of noise: white Gaussian noise and three from the MIT-BIH noise stress test [40] (e.g., baseline wander, electrode movement, and movement artifact). For each of these noise types, we added the noise in increasing amounts, up to 100% of the signal's power (i.e., SNR = 0 dB). Furthermore, we scaled the noise relative to each epoch's signal power to compensate for variations across the recording. These experiments verify that it takes substantial noise levels to affect the performance meaningfully (Fig. 10b). The exception is movement artifact noise, which causes a steeper roll-off. Moreover, with increasing movement artifact noise, the network increasingly scores epochs as Wake (Supplementary Fig. S4a).

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

Next, because we designed the network to use context, evaluating what happens when we drastically modify the context is helpful. Therefore, we modified the context by reversing or shuffling the epoch order (Fig. 10c). Reversing the epoch order significantly affects the performance negatively—even more so when we shuffle the epoch order. We also evaluated how losing entire epochs, possibly due to intermittent connections, affects the model's performance. For scale, we lost an average of four minutes per recording (and a maximum of 3.6 hr.) of the ECG data due to intermittent connections (Methods 2.3). To simulate additional data loss, we silenced (i.e., set the signal to a value of 0) individual epochs selected randomly from the recording. We analyzed the kappas of the intact and silenced epochs separately (Fig. 10e). Model performance decreases for both, indicating that the network uses the ECG input from surrounding epochs in determining the sleep stage for a given epoch (Fig. 10c)—and yet is robust enough to lose up to 35% and still exceed the "time-only" floor. In addition to silencing whole epochs, we also tested the effect of silencing portions of epochs (Supplementary Fig. S4b). We found that the center of the epoch was more informative for the model than the ends of the epoch.

Finally, we evaluated the usefulness of the two demographic inputs. First, we flipped the sex for all subjects. Next, we assigned an age with uniform probability from 5 to 90. The results show a significant effect when modifying age but not sex (Fig. 10f). The minuscule effect on sex performance mirrors the result showing no overall or stage-specific sex differences (Supplementary Fig. S2d). However, the larger decrease in performance due to age also showed up when stratifying the results by age (Supplementary Fig. S2a). Specifically, while the performance decrease with increasing age was slight for most stages, there was a pronounced decrease in N3 performance—similar to the N3 results above.

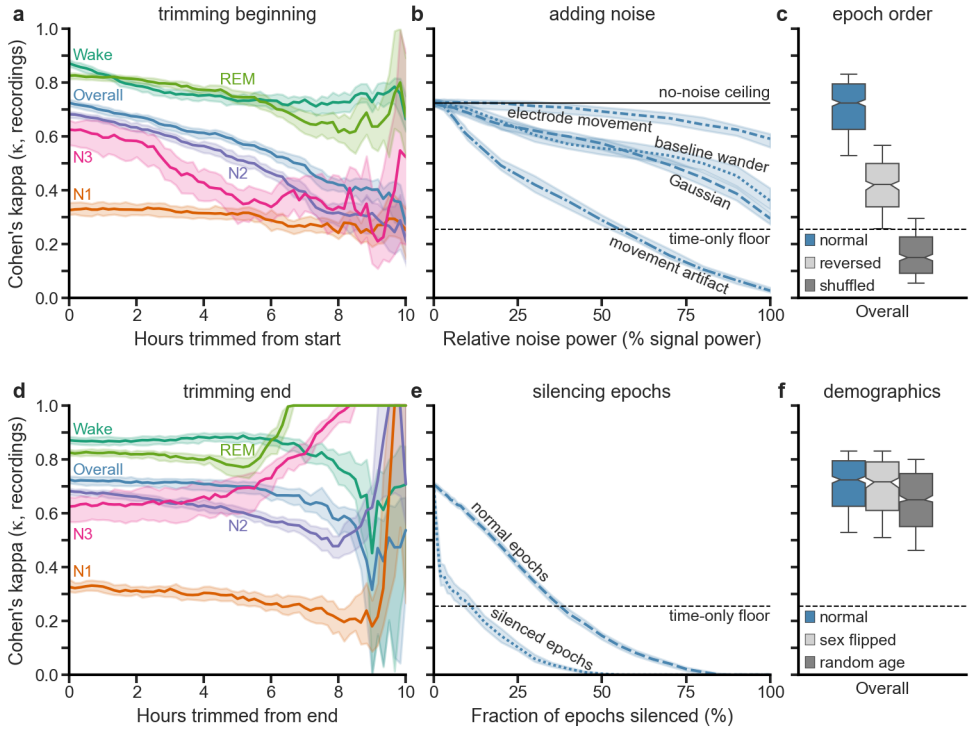


Fig. 10. Performance with noise or modifications to the testing set

(a) When trimming recordings from the start, there is a gradual downward trend in kappa for every stage except N1, indicating the importance of the early periods. (b) Likewise, performance decreases with relative

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

noise power (noise power divided by signal power) for four different noise sources (electrode movement, baseline wander, Gaussian, and movement artifact). Shown are a “no-noise ceiling” (zero noise added, black solid line) and a “time-only floor” (Fig. 7, black dashed line, which continues into panel c). The model exhibits remarkable robustness for all noise sources tested except movement artifact, as the performance exceeds the floor, even at 100% relative noise. (c) When we reverse or shuffle the epoch order of each recording, the performance is significantly worse, indicating that the model strongly relies on temporal order. (d) When trimming the recordings from the end, the performance is less affected than when trimmed from the start (panel a). This result indicates it will perform well even if the recording stopped prematurely. (e) Performance as a function of the fraction of all epochs silenced (replaced with zeros) quickly trends downward. Since the model uses context for intact and silenced epochs, it suggests that the disconnection of electrodes for extended periods will cause the model to suffer. (f) There is no change in performance when we flip the subject’s sex, suggesting some insensitivity to sex. In contrast, performance is significantly lower when we assign a random age. Shaded areas represent 95% CIs. Whiskers at P10 and P90.

3.5. Real-time scoring

Because some of the possible use cases for this technique would require real-time scoring, we also tested a slightly modified network variant. Note that our primary model scores the entire night of sleep at once, and information from past and future epochs contributes to classifying each epoch. Similar to the results from when we silenced random epochs (Fig. 10e), computing the normalized relative importance (Methods 2.12) of epochs showed that our primary network uses contextual information before and after to score each epoch (Supplementary Fig. S2c). Of note are the “blips” at powers of two that reflect the underlying structure of the network (Fig. 2).

We altered the network's structure to only use information from the current and past epochs (Fig. 3). Therefore, this real-time model only operates causally. When comparing our primary model with the real-time model on the entire testing set (Fig. 11), we see that the real-time model's performance is usually slightly lower than our primary model’s. However, the real-time performance is still non-inferior ($p < 0.025$). The single exception is still stage N3, which is slightly below the threshold—and, therefore, not non-inferior.

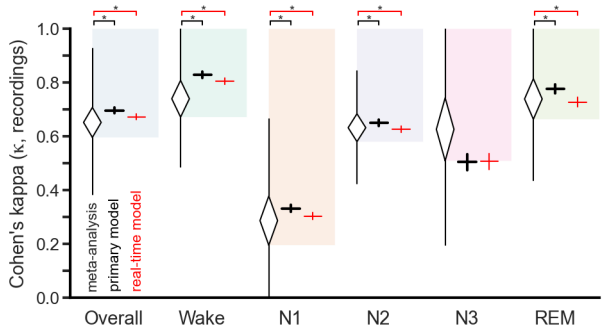


Fig. 11. Performance of real-time model variant

For overall and each stage, we duplicated the meta-analysis estimate and our primary model’s result on all five datasets to aid the reader (from Fig. 4 and Fig. 5). Then, for each stage, we included our real-time model’s results. The colored area is the non-inferior region. Almost all non-inferior comparisons using a t-test are significant (i.e., non-inferior; *p<0.025). The exception (not non-inferior) is the real-time model on N3. We tabulated the meta-analysis results in Supplementary Table S3 and the adjusted p-values in Supplementary Table S4.

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

When scoring in real time, we also assessed the performance across time to see if there were any changes due to the lack of future context. We found that N1 and REM (Supplementary Fig. S5b) performance took longer to reach their plateaus than our primary model (Supplementary Fig. S5a). However, the performance across time for each stage was generally relatively constant. The exception here is N3, where the performance rapidly drops much later in the night (Supplementary Fig. S3d).

3.6. Miscellaneous results

Numerous other supplementary results did not neatly fit into the previous sections but supported the main findings or covered related details, such as our loss function's performance. We present them in this section.

One significant finding was that kappa decreased as the sleep stage ratio approached all or nothing (i.e., the proportions deviated significantly from equality) (Supplementary Fig. S2b). This finding matches the expectation for kappa, as discussed in Methods 2.6.

Although we only selected 4,000 recordings, there were an additional 1,718 that still met the quality criteria (Methods 2.14). The results of those additional recordings mirror those of the recordings from the same source datasets in the test set (Supplementary Fig. S6a). We also wanted to compare the testing set results with an entirely held-out study to assess for any study-specific learning that had occurred. When controlling for age and sex, we found no performance decrease on data from unseen sources (Supplementary Fig. S6b). Finally, two tables summarize the kappas (Supplementary Table S7) and other possible classification metrics (Supplementary Table S8). However, we stress that the other classification metrics are not appropriate for inter-rater agreement, nor are they reported in the human PSG sleep literature.

Finally, we compare our loss function with others in Supplementary Table S9. Furthermore, there is an additional investigation into assessing if our loss function performs better on five-stage (i.e., default) or one-stage (i.e., one-vs-all) scoring in Supplementary Table S10.

4. Discussion

To the best of our knowledge, our study represents the first successful demonstration of five-stage sleep staging on par with expert-scored PSG without the aid of EEG. Similar EEG-less studies have shown correlations between sleep stages defined by PSG and data obtained from other non-EEG sensors. The advantages of replacing human scorers with automated algorithms are significantly reduced labor costs and increased inter-rater agreement. Additionally, using ECG has further advantages, such as a more user-friendly setup, a more robust signal, and broader accessibility to the scientific community and citizen scientists. Prior to our work, the suboptimal performance of EEG-less methods in five-stage classification had suggested that EEG would always be necessary for achieving clinically relevant sleep staging. However, our findings establish that ECG-based automated sleep staging can achieve comparable performance to PSG-based human sleep staging, thereby challenging the notion that EEG is indispensable.

4.1. On par with human-scored PSG

Regarding the performance of our model, it is noteworthy that it achieves expert-human level agreement overall (Fig. 4) and for each of the stages except N3 (Fig. 5). The N3 result discrepancy is interesting, as it only occurs for two of the five source datasets: MESA and WSC—the results

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

from the other three datasets (i.e., CCSHS, CFS, and CHAT) all achieve non-inferior (i.e., on-par) performance (Fig. 6). The drop in N3 kappa could be partially attributed to the worse N3 performance with age (Supplementary Fig. S2a), observed for human-scored PSG as well [43], which, in turn, is partly attributed to the fact that at older ages, the proportion of N3 drops precipitously. Half of the subjects in the CFS dataset were also older than the average age; when investigating the kappa stratified by decade and disaggregated by source dataset (Supplementary Fig. S3c), we notice that the older CFS subjects also show this decrease in performance. However, note that the proportion of N3 for MESA and WSC in the older subjects is even lower than expected for that range of ages [57] (Supplementary Fig. S3a,b; MESA and WSC shaded regions as compared to the dotted black line). As previously mentioned, the kappa will tend towards zero as the proportion of a stage approaches zero. Therefore, we expect that even slight disagreements in N3 scores—when their proportion is nearly zero—would lead to outsized decreases in N3 kappa. Finally, we should mention that MESA was studying atherosclerosis—which, by definition, negatively impacts cardiac physiology. These facts lead to the conclusion that there might be something different about the subjects in MESA and WSC. It is also possible that this difference could be situational and only reflect their ability to enter N3 on those particular nights. On the other hand, it does bear to mention that both the human (Supplementary Fig. S3a) and model (Supplementary Fig. S3b) essentially agree about the N3 proportion. This agreement suggests that our model is not unusually performing worse in classifying N3 in older subjects with lower than expected (based on age) near-zero proportions of N3.

4.2. Significantly better than other EEG-less models

Our model significantly outperforms other EEG-less models, irrespective of five-, four-, or three-stage scoring (Fig. 7). While the current literature on EEG-less methods (including current commercial sleep-tracking devices) is more extensive than what we referenced here, we excluded papers for one or more reasons. The possible reasons include not listing kappa, only considering two-stage scoring (i.e., Wake/Sleep), or their evaluation set was not independent of their training set (Supplementary Discussion 6.3.2 for details). It bears stressing that none of the papers that we excluded for methodological reasons (and reported a kappa) had kappas that were higher than those results that we included for the same stage granularity.

4.3. Concordant and robust scoring

Moreover, the consistency between the model's classifications and the human-scored stages is evident when analyzing the row-normalized contingency table (Fig. 8a), stage transition rates (Fig. 9a), and stage transition matrix (Fig. 9b,c). These findings strongly support the argument that our model scores sleep stages similarly to and on par with human-scored PSG.

Like human scorers, the network also incorporates contextual information (Fig. 10c,e and Supplementary Fig. S2c). Furthermore, the performance across time demonstrates that, for most stages, the model's performance is relatively insensitive to time during the recording (Supplementary Fig. S5a). Although this is less generally true at the ends, and perhaps with N3 and REM. Notably, the network begins predominantly scoring epochs classified as N3 by human scorers as N2 after approximately 9 hours (Supplementary Fig. S3d). This finding indicates at least two possibilities. First, there is a potential qualitative difference between N3 sleep occurring at the

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

night's end versus earlier. Second, and more likely, is that given the low proportion of N3 later in the night, the performance drop is an artifact.

Finally, our model demonstrates robustness to added noise and data corruption, which is advantageous in the variable and harsher environments outside the clinic. Specifically, the model was able to perform very well with common and harsh noise sources—even with relative noise powers exceeding 50% (Fig. 10b). Additionally, although artificial, findings show that the model is less affected by noise at both ends of the epoch (Supplementary Fig. S4b). Moreover, the results indicate that our model performs best when the subject has the relatively non-intrusive ECG strap attached at least 30 minutes before falling asleep (Supplementary Fig. S2e). However, the performance is not contingent upon knowing when sleep began or ended.

4.4. Real-time scoring is possible

Another advantage our method has over PSG is the ability to score in real time—at the same performance level (Fig. 11). While the performance at the beginning of the night (Supplementary Fig. S5b) is slightly less than the same period on our primary model (Supplementary Fig. S5a), this result could be a combination of two issues. First, kappa tends toward zero when the proportion of one or more stages approaches zero [23]. Second, the real-time model might need sufficient initial context before it can score at its best level. Regardless, having the ability to score in real time opens additional interventions and applications.

4.5. Limitations

4.5.1. No standardized benchmark

We must address the most significant issue in sleep staging literature—common to both human-scored PSG and machine learning using PSG or EEG-less inputs: There is currently no standardized benchmark (i.e., a single dataset used to evaluate humans and models). Notwithstanding, training materials for new human scorers exist, and our meta-analysis reveals that numerous studies have compared humans on small but disparate datasets. However, every new study (often even by the same lab) will use different recordings. This issue leads to complications when comparing inter-rater agreements. The agreement could be worse not only from expected ambiguity stemming from the scoring rules' flexibility but also from differences in equipment and scorer training. Moreover, it has the knock-on effect that two studies using two different source datasets scored by two (or more) different scorers could report substantially different kappas. Furthermore, hypothetically, if tested on a single altogether different dataset, the two studies could perform the same, or the worse one could perform better.

Additionally, although there are now publicly available datasets, the researchers often use different portions of the same dataset. Unfortunately, while studies might report using the same dataset (e.g., CFS), this is not enough for a standard. Studies often have their own (unreported) quality metrics for which they will exclude some recordings (i.e., when comparing sample sizes versus the source dataset, they often do not match). Above all, researchers rarely report which specific recordings they included in the training, evaluation, and testing sets. That said, the meta-analysis and comparisons with EEG-less results have some semblance of consistency—considering stage granularity and other differences.

We tried to account for the dataset diversity issue by taking data from five different source datasets to increase the diversity of equipment, subjects, and human training and agreement. Furthermore, from the outset of our study, we knew one of our end goals was the creation of a

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

standardized dataset. This goal is why we meticulously documented and developed a pipeline for automatic curation and random selection of the data. This goal also has the vital side-effect of substantially reducing, if not eliminating, the researcher's sampling bias. Additionally, we will provide the filenames and details for each of the sets so that others can train, evaluate, and test on exactly the same sets in the future. We hope this will go some way towards providing future researchers in the field with a standardized benchmark for future comparisons with our current state-of-the-art results and others.

4.5.2. Potential biases in sampling

As evidenced by our results and discussion regarding the much lower-than-expected N3 ratios in the MESA and WSC datasets (see 4.1 above), there is significant diversity in the sex, race, ethnicity, and medical conditions of the subjects in our source datasets. For instance, while most of the subjects had no reported medical condition, some conditions included diabetes, Alzheimer's, coronary artery disease, and depression. One can also see this diversity in the meta-analysis input studies. However, the effects of this diversity are not always intuitive, such as sometimes seeing higher kappas with controls than patients. Apart from the significant differences for MESA and WSC, no other dataset showed such consistent performance differences from the mean. Because we did not analyze demographic data other than age or sex, we do not know whether other variables, such as ethnicity or medical condition, would show noticeable performance effects.

In addition to the diversity of subjects of our source datasets, there were numerous differences in equipment and aims between those datasets. These differences were partially due to the lower importance of ECG during PSG and included electrode placement, sampling rates, and quantization. In addition to harmonizing these differences where possible (which can also cause issues), we discarded numerous recordings because of poor data quality (Methods 2.2.3). Finally, because we resampled all data to 256 Hz for our model, we could not determine the effect of the sampling rate on performance. Unfortunately, because the studies were largely homogenous in their equipment and sampling rates, we cannot disambiguate performance that might be due to equipment from the study-specific differences we already highlighted.

4.5.3. Network robustness

Because it is still unknown what exactly characterizes the ECG of a given sleep stage, it is not yet possible to generate data synthetically. This limited our ability to explore the robustness of the network to only adding noise or modifying the recordings (e.g., silencing epochs), as opposed to synthesizing atypical heartbeats or rhythms. Relatedly, along with the lack of "ground truth" scores, we could not explore purposeful misclassifications by modifying an epoch to "look" more like another stage. A possible exception is with movement artifact noise (Fig. 10b), which had the effect of eventually making all epochs look like Wake (Supplementary Fig. S4a).

4.5.4. Does age and sex matter?

We designed the model to take the age and sex of the subject as input. However, in some cases, this demographic information may not be known because of missing information or privacy concerns. There was no difference seen in the performance based on sex (Supplementary Fig. S2d) or when we "flipped" the sex (Fig. 10f). But performance slightly decreased when we assigned a random age (Fig. 10f). However, we did not assess the ultimate performance impact of just removing this input from the model altogether. Given that the modified sex had no performance impact, it is possible that training a model without either input would not significantly affect the model's top-line performance.

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

4.5.5. Black box nature of neural networks

Moreover, there are consequences to using neural networks. We evolved our model from simpler machine learning techniques to the current deep neural network (see Supplementary Methods 6.1.1 for a history) to improve the performance. This performance improved at the expense of explainability (i.e., simpler models' decisions are explainable; neural networks, generally, are not). While the nascent field of "Explainable AI" is developing techniques to explain the decisions of these black boxes, they have a long way to go. For research applications, few would consider this an insurmountable problem. However, for clinical applications, it may cause some hesitation. The E.U. just passed a broad law on AI use—including in high-risk settings such as medical applications—and somewhat similar bills have been considered in the U.S. Regardless, there are still currently no laws or regulations in the U.S. or E.U. that require a model to be able to explain its decisions, even in medical applications. We feel that future laws and regulations should address this gap more directly, even if it will highlight issues with the current human clinical judgments (e.g., biases and motivated reasoning). We hope that the techniques for explaining networks and determining their limits will address the explainability gap and meet the standards of any possible future rules.

4.5.6. Foundational assumptions

As mentioned, the data we used came from studies that scored their data using either R&K or AASM. Moreover, although there are slight differences in the criteria for the similarly named stages, it was necessary to harmonize them. The refinements in the scoring criteria lead to some differences—even when scorers use both methodologies for the same recording [58]. While the total sleep time (i.e., Wake vs. sleep) and REM scoring will be nearly identical, others have found that for the non-REM stages (i.e., N1/N2/N3), the distributions will slightly change (e.g., a decrease in N2 (~5%), going to N1 and N3). Because we used a single class label for each similarly named stage, the network likely had to make compromises where the two methodologies would disagree.

Finally, and foundationally, our work necessitates using Cohen's kappa and the R&K and AASM scoring rules. First, Cohen's kappa, the most commonly reported inter-rater agreement measure, is not without its detractors. As mentioned, one significant consequence is that as the proportion of a stage approaches zero, the kappa does as well (Supplementary Fig. S2b) [23]. While some have suggested modifications to kappa to account for bias and prevalence, others accept the limitations of a single value representing so much information [23]. However, researchers rarely use these modified versions, and we found no reporting of these "adjusted" kappas in the sleep stage literature. Secondly, the current practice and orthodoxy is to use R&K (now AASM) scoring rules, which assume that one can cleanly segment sleep into self-consistent discrete stages. However, many alternative methods have been used to score sleep, and critiques of these rules exist [59]. Furthermore, some data supports a more heterogeneous—both across and within a stage—understanding of sleep stages [60], e.g., N3 at the beginning of the night could differ in some aspects from N3 towards the end.

4.6. Cardiosomnography

The effectiveness of ECG in sleep medicine for detecting sleep disorders, such as apnea, has been established by numerous studies [13], [61], [62]. Building upon these findings, we propose that cardiosomnography—a sleep study conducted with ECG only—can complement and supplement PSG in sleep medicine. Compared to PSG, cardiosomnography offers advantages in terms of cost-effectiveness and simplicity.

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

The adoption of cardiosomnography has the potential to facilitate more accessible sleep medicine, thereby enabling further research, interventions, and applications. For instance, it has been found that in Alzheimer's disease, a feedback loop can arise where the disease's progression can lead to increasingly disturbed sleep [63]. In turn, chronic poor sleep can contribute to the accumulation of β-amyloid [64]. Continuous and inexpensive sleep monitoring and disease progression tracking for at-risk individuals could guide treatments and potentially improve outcomes.

Additionally, cardiosomnography could be a valuable tool for interventions. Research has indicated that playing pink noise only during N3 increases the amplitude of slow oscillation waves and the proportion of N3 during the night [4], [65]. Bringing this intervention out of the clinic with real-time ECG-based sleep staging could be helpful for those individuals with mild cognitive impairments, who typically have less N3 [66]. Furthermore, for consumer applications, it has been found that entering and quickly exiting N1 can enhance creativity [67]. With a less cumbersome method of monitoring and alerting users during N1, developers could make this intervention widely available.

Studying sleep before the dawn of modern society and its use of electronics, which researchers have long suspected of delaying and reducing sleep, has been a valuable tool for understanding the function of sleep. Sleep studies in pre-industrial societies have used actigraphy for its ease of use and cost [68]. With our single-lead-based ECG, one does not have to compromise on the quality and granularity of sleep staging at the same level of cost and convenience as before.

We feel that cardiosomnography holds exciting potential to benefit researchers, physicians, and entrepreneurs by more easily providing objective measures of sleep for interventions and sleep medicine as a whole. Moreover, it opens possibilities for as-yet-unknown applications and could expand the scope of sleep-related research, interventions, and healthcare.

5. Conclusion

In summary, this study introduces a groundbreaking approach to sleep stage classification that utilizes a single-lead ECG-based neural network. We have successfully demonstrated that our method achieves expert-level agreement with the gold-standard PSG—without the need for expensive and cumbersome equipment. This advancement challenges the traditional reliance on EEG for reliable sleep staging and paves the way for more accessible, cost-effective sleep studies. By enabling access to high-quality sleep analysis outside clinical settings, our research holds the potential to expand the reach of sleep medicine significantly. This availability could improve health outcomes by better understanding and monitoring of sleep patterns. Our findings underscore the viability of cardiosomnography (i.e., ECG-based sleep studies) as a standalone tool for a sleep study, potentially signifying a significant advancement in sleep research and healthcare interventions.

Conflicts of interest

There are no conflicts of interest that could inappropriately influence this research work.

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

Data availability

All human input data for this study came from the datasets listed and are available at the National Sleep Research Resource (https://sleepdata.org/).

Code availability

We have provided all code, model weights, and set filenames to reproduce everything in the Methods and Results on GitHub at (https://github.com/adammj/ecg-sleep-staging/).

Author contributions

A.M.J. and B.R.S. conceived the study. A.M.J. refined and crystalized the study, performed the literature and dataset searches, developed the algorithms for data pre-processing and selection, designed and trained the deep network, created the loss function, analyzed the data, produced the results, and wrote the manuscript. B.R.S. regularly assisted, including with discussing results and revising the manuscript. L.I. provided critical feedback on the results and text. All authors discussed the results and contributed to the final manuscript.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgments

The authors acknowledge the use of the Opuntia, Sabine, and Carya clusters and the support from the Research Computing Data Core at the University of Houston to carry out the research presented here.

The Cleveland Children's Sleep and Health Study (CCSHS) was supported by grants from the National Institutes of Health (RO1HL60957, K23 HL04426, RO1 NR02707, M01 Rrmpd0380-39). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

The Cleveland Family Study (CFS) was supported by grants from the National Institutes of Health (HL46380, M01 RR00080-39, T32-HL07567, RO1-46380). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

The Childhood Adenotonsillectomy Trial (CHAT) was supported by the National Institutes of Health (HL083075, HL083129, UL1-RR-024134, UL1 RR024989). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

The Multi-Ethnic Study of Atherosclerosis (MESA) Sleep Ancillary study was funded by NIH-NHLBI Association of Sleep Disorders with Cardiovascular Health Across Ethnic Groups (RO1 HL098433). MESA is supported by NHLBI funded contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

95165, N01-HC-95166, N01-HC-95167, N01-HC-95168 and N01-HC-95169 from the National Heart, Lung, and Blood Institute, and by cooperative agreements UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420 funded by NCATS. The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

The National Heart, Lung, and Blood Institute provided funding for the ancillary MrOS Sleep Study (MROS), "Outcomes of Sleep Disorders in Older Men," under the following grant numbers: R01 HL071194, R01 HL070848, R01 HL070847, R01 HL070842, R01 HL070841, R01 HL070837, R01 HL070838, and R01 HL070839. The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

This Wisconsin Sleep Cohort Study (WSC) was supported by the U.S. National Institutes of Health, National Heart, Lung, and Blood Institute (R01HL62252), National Institute on Aging (R01AG036838, R01AG058680), and the National Center for Research Resources (1UL1RR025011). The National Sleep Research Resource was supported by the U.S. National Institutes of Health, National Heart Lung and Blood Institute (R24 HL114473, 75N92019R002).

References

[1] J. A. Hobson, “Sleep is of the brain, by the brain and for the brain,” *Nature*, vol. 437, no. 7063, pp. 1254–1256, 2005, doi: 10/fhvs9s.

[2] A. Rechtschaffen and A. Kales, *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects.* Los Angeles: Brain Information Service/Brain Research Institute, University of California, 1968.

[3] C. Iber, S. Ancoli-Israel, A. L. Chesson Jr., and S. F. Quan, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules Terminology and Technical Specifications 1st ed.* 2007.

[4] M. M. Schade, G. M. Mathew, D. M. Roberts, D. Gartenberg, and O. Buxton, “Enhancing Slow Oscillations and Increasing N3 Sleep Proportion with Supervised, Non-Phase-Locked Pink Noise and Other Non-Standard Auditory Stimulation During NREM Sleep,” *Nat. Sci. Sleep*, vol. Volume 12, pp. 411–429, Jul. 2020, doi: 10/gk6bh5.

[5] M. R. Zielinski, J. T. McKenna, R. W. McCarley, and 1 Veterans Affairs Boston Healthcare System, West Roxbury, MA 02132, USA and Harvard Medical School, Department of Psychiatry, “Functions and Mechanisms of Sleep,” *AIMS Neurosci.*, vol. 3, no. 1, pp. 67–104, 2016, doi: 10/ghh8mp.

[6] R. S. Rosenberg and S. Van Hout, “The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring,” *J. Clin. Sleep Med.*, vol. 09, no. 01, pp. 81–87, Jan. 2013, doi: 10/gbcf3r.

[7] Y. J. Lee, J. Y. Lee, J. H. Cho, and J. H. Choi, “Interrater reliability of sleep stage scoring: a meta-analysis,” *J. Clin. Sleep Med.*, vol. 18, no. 1, pp. 193–202, Jan. 2022, doi: 10/gtp3jw.

[8] L. Fiorillo *et al.*, “Automated sleep scoring: A review of the latest approaches,” *Sleep Med. Rev.*, vol. 48, p. 101204, Dec. 2019, doi: 10/ggrt3g.

[9] H. W. Loh *et al.*, “Automated Detection of Sleep Stages Using Deep Learning Techniques: A Systematic Review of the Last Decade (2010–2020),” *Appl. Sci.*, vol. 10, no. 24, p. 8963, Dec. 2020, doi: 10/gnvdnf.

[10] D. C. Lim *et al.*, “Reinventing polysomnography in the age of precision medicine,” *Sleep Med. Rev.*, vol. 52, p. 101313, Aug. 2020, doi: 10/gtp3jz.

[11] I. Perez-Pozuelo *et al.*, “The future of sleep health: a data-driven revolution in sleep science and medicine,” *Npj Digit. Med.*, vol. 3, no. 1, p. 42, Dec. 2020, doi: 10/gk4xzs.

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

[12] T. Poppa and A. Bechara, “The somatic marker hypothesis: revisiting the role of the ‘body-loop’ in decision-making,” *Curr. Opin. Behav. Sci.*, vol. 19, pp. 61–66, Feb. 2018, doi: 10/gmw899.
[13] T. Wang, J. Yang, Y. Song, F. Pang, X. Guo, and Y. Luo, “Interactions of central and autonomic nervous systems in patients with sleep apnea–hypopnea syndrome during sleep,” *Sleep Breath.*, Jul. 2021, doi: 10/gn2gpf.
[14] G.-Q. Zhang *et al.*, “The National Sleep Research Resource: towards a sleep data commons,” *J. Am. Med. Inform. Assoc.*, vol. 25, no. 10, pp. 1351–1358, Oct. 2018, doi: 10/gdntnm.
[15] C. L. Rosen *et al.*, “Prevalence and risk factors for sleep-disordered breathing in 8- to 11-year-old children: association with race and prematurity,” *J. Pediatr.*, vol. 142, no. 4, pp. 383–389, Apr. 2003, doi: 10/fmsgn8.
[16] S. Redline *et al.*, “The Familial Aggregation of Obstructive Sleep Apnea,” *Am. J. Respir. Crit. Care Med.*, vol. 151, no. 3 Pt 1, pp. 682–687, Mar. 1995, doi: 10/ghv3pb.
[17] C. L. Marcus *et al.*, “A Randomized Trial of Adenotonsillectomy for Childhood Sleep Apnea,” *N. Engl. J. Med.*, vol. 368, no. 25, pp. 2366–2376, Jun. 2013, doi: 10/ggv7j4.
[18] X. Chen *et al.*, “Racial/Ethnic Differences in Sleep Disturbances: The Multi-Ethnic Study of Atherosclerosis (MESA),” *SLEEP*, Jun. 2015, doi: 10/gftdrx.
[19] T. Young, M. Palta, J. Dempsey, P. E. Peppard, F. J. Nieto, and K. M. Hla, “Burden of sleep apnea: rationale, design, and major findings of the Wisconsin Sleep Cohort study,” *WMJ Off. Publ. State Med. Soc. Wis.*, vol. 108, no. 5, pp. 246–249, Aug. 2009.
[20] A. C. Charles, C. Z. Janet, M. R. Joseph, C. M.-E. Martin, and D. W. Elliot, “Timing of REM sleep is coupled to the circadian rhythm of body temperature in man,” *Sleep*, vol. 2, no. 3, pp. 329–346, 1980, doi: 10/gkhp74.
[21] S. Bai, J. Z. Kolter, and V. Koltun, “An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling,” *ArXiv180301271 Cs*, Apr. 2018, doi: 10/gtp3jb.
[22] Z. Xie, I. Sato, and M. Sugiyama, “Understanding and Scheduling Weight Decay,” *ArXiv201111152 Cs*, Sep. 2021, doi: 10/gtp3jc.
[23] W. Vach, “The dependence of Cohen’s kappa on the prevalence does not matter,” *J. Clin. Epidemiol.*, vol. 58, no. 7, pp. 655–661, Jul. 2005, doi: 10/d5k8wx.
[24] M. J. Warrens, “Cohen’s kappa is a weighted average,” Jun. 2011, doi: 10/c92kt8.
[25] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. in Springer Series in Statistics. New York, NY: Springer New York, 2009. doi: 10.1007/b94608.
[26] J. M. Johnson and T. M. Khoshgoftaar, “Survey on deep learning with class imbalance,” *J. Big Data*, vol. 6, no. 1, p. 27, Dec. 2019, doi: 10/gghpn6.
[27] St. Kubicki, L. Höller, I. Berg, C. Pastelak-Price, and R. Dorow, “Sleep EEG Evaluation: A Comparison of Results Obtained by Visual Scoring and Automatic Analysis with the Oxford Sleep Stager,” *Sleep*, vol. 12, no. 2, pp. 140–149, Mar. 1989, doi: 10/gtp3j5.
[28] N. Schaltenbrand *et al.*, “Sleep Stage Scoring Using the Neural Network Model: Comparison Between Visual and Automatic Analysis in Normal Subjects and Patients,” *Sleep*, vol. 19, no. 1, pp. 26–35, Jan. 1996, doi: 10/gtp3j4.
[29] C. W. Whitney *et al.*, “Reliability of Scoring Respiratory Disturbance Indices and Sleep Staging,” *Sleep*, vol. 21, no. 7, pp. 749–757, Oct. 1998, doi: 10/gtp3j3.
[30] S. D. Pittman *et al.*, “Assessment of Automated Scoring of Polysomnographic Recordings in a Population with Suspected Sleep-disordered Breathing,” vol. 27, no. 7, 2004.
[31] J. L. Fleiss, B. A. Levin, and M. Cho. Paik, *Statistical methods for rates and proportions.*, 3rd ed. / Joseph 1. Fleiss, Bruce Levin, Myunghee Cho Paik. Hoboken, N.J: Wiley-Interscience, 2003. doi: 10.1002/0471445428.

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

[32] D. Luo, X. Wan, J. Liu, and T. Tong, “Optimally estimating the sample mean from the sample size, median, mid-range, and/or mid-quartile range,” *Stat. Methods Med. Res.*, vol. 27, no. 6, pp. 1785–1805, Jun. 2018, doi: 10/gdn32t.
[33] X. Wan, W. Wang, J. Liu, and T. Tong, “Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range,” *BMC Med. Res. Methodol.*, vol. 14, no. 1, p. 135, Dec. 2014, doi: 10/f7xmgr.
[34] R. DerSimonian and N. Laird, “Meta-analysis in clinical trials,” *Control. Clin. Trials*, vol. 7, no. 3, pp. 177–188, Sep. 1986, doi: 10/fdpm2j.
[35] K. Nagashima, H. Noma, and T. A. Furukawa, “Prediction intervals for random-effects meta-analysis: A confidence distribution approach,” *Stat. Methods Med. Res.*, vol. 28, no. 6, pp. 1689–1702, Jun. 2019, doi: 10/ggmbqv.
[36] S. G. Thompson and S. J. Sharp, “Explaining heterogeneity in meta-analysis: a comparison of methods,” *Stat. Med.*, vol. 18, no. 20, pp. 2693–2708, Oct. 1999, doi: 10/bpw8n4.
[37] M. D. Rothmann, B. L. Wiens, and I. S. F. Chan, *Design and Analysis of Non-Inferiority Trials*, 0 ed. Chapman and Hall/CRC, 2011. doi: 10.1201/b11039.
[38] Y. D. Alqurashi *et al.*, “A novel in-ear sensor to determine sleep latency during the Multiple Sleep Latency Test in healthy adults with and without sleep restriction,” *Nat. Sci. Sleep*, vol. Volume 10, pp. 385–396, Nov. 2018, doi: 10/ggzc2g.
[39] Y. Hochberg, “A sharper Bonferroni procedure for multiple tests of significance,” *Biometrika*, vol. 75, no. 4, pp. 800–802, 1988, doi: 10/c2v8s5.
[40] G. B. Moody, W. Muldrow, and R. G. Mark, “A noise stress test for arrhythmia detectors,” *Comput. Cardiol.*, vol. 11, no. 3, pp. 381–384, 1984.
[41] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic Attribution for Deep Networks,” *ArXiv170301365 Cs*, Jun. 2017, doi: 10/grx4kq.
[42] T. Blackwell *et al.*, “Associations Between Sleep Architecture and Sleep-Disordered Breathing and Cognition in Older Community-Dwelling Men: The Osteoporotic Fractures in Men Sleep Study,” *J. Am. Geriatr. Soc.*, vol. 59, no. 12, pp. 2217–2225, Dec. 2011, doi: 10/c86vbm.
[43] D. Kunz *et al.*, “Interrater Reliability Between Eight European Sleep-Labs In Healthy Subjects Of All Age Groups,” *Biomed. Tech. Eng.*, vol. 45, no. s1, pp. 433–434, Jan. 2000, doi: 10/c9jzd6.
[44] G. Brandenberger *et al.*, “Age-related changes in cardiac autonomic control during sleep.,” *J. Sleep Res.*, vol. 12, no. 3, pp. 173–80, 2003, doi: 10/dtpdb6.
[45] H. Sun *et al.*, “Sleep staging from electrocardiography and respiration with deep learning,” *Sleep*, vol. 43, no. 7, p. zsz306, Jul. 2020, doi: 10/gk5z72.
[46] M. Radha *et al.*, “A deep transfer learning approach for wearable sleep stage classification with photoplethysmography,” *Npj Digit. Med.*, vol. 4, no. 1, p. 135, Sep. 2021, doi: 10/gnkxbz.
[47] B. M. Wulterkens *et al.*, “It is All in the Wrist: Wearable Sleep Staging in a Clinical Population versus Reference Polysomnography,” *Nat. Sci. Sleep*, vol. Volume 13, pp. 885–897, Jun. 2021, doi: 10/gtp3jx.
[48] P. Fonseca *et al.*, “Automatic sleep staging using heart rate variability, body movements, and recurrent neural networks in a sleep disordered population,” *Sleep*, vol. 43, no. 9, p. zsaa048, Sep. 2020, doi: 10/gg22z9.
[49] N. Sridhar *et al.*, “Deep learning for automated sleep staging using instantaneous heart rate,” *Npj Digit. Med.*, vol. 3, no. 1, p. 106, Dec. 2020, doi: 10/gmxn9x.
[50] Z. Beattie, A. Pantelopoulos, A. Ghoreyshi, Y. Oyang, A. Statan, and C. Heneghan, “Estimation of Sleep Stages Using Cardiac and Accelerometer Data from a Wrist-Worn Device,” *Sleep*, vol. 40, p. A26, 2017, doi: 10/gnr56x.

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

[51] H. Yoon, S. H. Hwang, J.-W. Choi, Y. J. Lee, D.-U. Jeong, and K. S. Park, “REM sleep estimation based on autonomic dynamics using R–R intervals,” *Physiol. Meas.*, vol. 38, no. 4, pp. 631–651, Apr. 2017, doi: 10/gnxqqx.
[52] A. Domingues, T. Paiva, and J. M. Sanches, “Hypnogram and sleep parameter computation from activity and cardiovascular data,” *IEEE Trans. Biomed. Eng.*, vol. 61, no. 6, pp. 1711–1719, 2014, doi: 10/f563cd.
[53] T. Willemen *et al.*, “An Evaluation of Cardiorespiratory and Movement Features With Respect to Sleep-Stage Classification,” *IEEE J. Biomed. Health Inform.*, vol. 18, no. 2, pp. 661–669, Mar. 2014, doi: 10/gnxqrf.
[54] C. C. R. Sady, U. S. Freitas, A. Portmann, J.-F. Muir, C. Letellier, and L. A. Aguirre, “Automatic sleep staging from ventilator signals in non-invasive ventilation,” *Comput. Biol. Med.*, vol. 43, no. 7, pp. 833–839, Aug. 2013, doi: 10/gphf3q.
[55] H. Danker-Hopfe *et al.*, “Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard,” *J. Sleep Res.*, vol. 18, no. 1, pp. 74–84, 2009, doi: 10/bcnw9b.
[56] A. Laffan, B. Caffo, B. J. Swihart, and N. M. Punjabi, “Utility of Sleep Stage Transitions in Assessing Sleep Continuity,” *Sleep*, vol. 33, no. 12, 2010, doi: 10/gtp3kp.
[57] H. Danker-Hopfe *et al.*, “Percentile Reference Charts for Selected Sleep Parameters for 20- to 80-Year-Old Healthy Subjects from the SIESTA Database. Referenzkurven fur ausgewahlte Schlafparameter 20- bis 80-jahriger gesunder Personen aus der SIESTA-Datenbank,” *Somnologie*, vol. 9, no. 1, pp. 3–14, Feb. 2005, doi: 10/fw7r97.
[58] D. Moser *et al.*, “Sleep Classification According to AASM and Rechtschaffen & Kales: Effects on Sleep Scoring Parameters,” vol. 32, no. 2, p. 11, 2009, doi: 10/gnsptz.
[59] H. Schulz, “Rethinking Sleep Analysis: Comment on the AASM Manual for the Scoring of Sleep and Associated Events,” *J. Clin. Sleep Med.*, vol. 04, no. 02, pp. 99–103, Apr. 2008, doi: 10/gtp3j2.
[60] C. Metzner, A. Schilling, M. Traxdorf, H. Schulze, and P. Krauss, “Sleep as a random walk: a super-statistical analysis of EEG data across sleep stages,” *Commun. Biol.*, vol. 4, no. 1, p. 1385, Dec. 2021, doi: 10/gsxkbg.
[61] T. Penzel, J. McNames, P. de Chazal, B. Raymond, A. Murray, and G. Moody, “Systematic comparison of different algorithms for apnoea detection based on electrocardiogram recordings,” *Med. Biol. Eng. Comput.*, vol. 40, no. 4, pp. 402–407, Jul. 2002, doi: 10/d2p6zw.
[62] H. Hilmisson, N. Lange, and S. P. Duntley, “Sleep apnea detection: accuracy of using automated ECG analysis compared to manually scored polysomnography (apnea hypopnea index),” *Sleep Breath.*, vol. 23, no. 1, pp. 125–133, Mar. 2019, doi: 10/ggmn7c.
[63] Y.-E. S. Ju, B. P. Lucey, and D. M. Holtzman, “Sleep and Alzheimer disease pathology—a bidirectional relationship,” *Nat. Rev. Neurol.*, vol. 10, no. 2, pp. 115–119, Feb. 2014, doi: 10/ggxcff.
[64] B. M. Brown *et al.*, “The Relationship between Sleep Quality and Brain Amyloid Burden,” *Sleep*, vol. 39, no. 5, pp. 1063–1068, May 2016, doi: 10/f8j9d9.
[65] M. Navarrete, J. Schneider, H.-V. V. Ngo, M. Valderrama, A. J. Casson, and P. A. Lewis, “Examining the optimal timing for closed-loop auditory stimulation of slow-wave sleep in young and older adults,” *Sleep*, vol. 43, no. 6, p. zsz315, Jun. 2020, doi: 10/grr9bf.
[66] C. E. Westerberg *et al.*, “Concurrent Impairments in Sleep and Memory in Amnestic Mild Cognitive Impairment,” *J. Int. Neuropsychol. Soc.*, vol. 18, no. 03, pp. 490–500, May 2012, doi: 10/f4kzjm.
[67] C. Lacaux *et al.*, “Sleep onset is a creative sweet spot,” *Sci. Adv.*, vol. 7, no. 50, p. eabj5866, Dec. 2021, doi: 10/g8sr.

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

[68] G. Yetish *et al.*, “Natural Sleep and Its Seasonal Variations in Three Pre-industrial Societies,” *Curr. Biol.*, vol. 25, no. 21, pp. 2862–2868, Nov. 2015, doi: 10/f7w8d7.
[69] R. G. Norman, I. Pal, C. Stewart, J. a Walsleben, and D. M. Rapoport, “Interobserver agreement among sleep scorers from different centers in a large dataset.,” *Sleep*, vol. 23, no. 7, pp. 901–908, 2000, doi: 10/c9r7hf.
[70] W. R. Ruehland *et al.*, “The 2007 AASM Recommendations for EEG Electrode Placement in Polysomnography: Impact on Sleep and Cortical Arousal Scoring,” *Sleep*, vol. 34, no. 1, pp. 73–81, Jan. 2011, doi: 10/gpb2mm.
[71] T. Penzel *et al.*, “Reliablität der visuellen schlafauswertung nach rechtschaffen und kales von acht aufzeichnungen durch neun schlaflabore: Reliability of visual evaluation of sleep stages according to rechtschaffen and kales from eight polysomnographs by nine sleep centres,” *Somnologie*, vol. 7, no. 2, pp. 49–58, 2003, doi: 10/b3xwjb.
[72] H. Danker-Hopfe *et al.*, “Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders: IRR of sleep stage scoring in patients,” *J. Sleep Res.*, vol. 13, no. 1, pp. 63–69, Mar. 2004, doi: 10/bs3kmk.
[73] J. R. Shambroom, S. E. Fábregas, and J. Johnstone, “Validation of an automated wireless system to monitor sleep in healthy adults,” *J. Sleep Res.*, vol. 21, no. 2, pp. 221–230, Apr. 2012, doi: 10/c9tqkt.
[74] R. Elliott, S. McKinley, P. Cistulli, and M. Fien, “Characterisation of sleep in intensive care using 24-hour polysomnography: an observational study,” *Crit. Care*, vol. 17, no. 2, p. R46, 2013, doi: 10/gj47gj.
[75] X. Zhang *et al.*, “Process and outcome for international reliability in sleep scoring,” *Sleep Breath.*, vol. 19, no. 1, pp. 191–195, Mar. 2015, doi: 10/f63cv7.
[76] S. Deng *et al.*, “Interrater agreement between American and Chinese sleep centers according to the 2014 AASM standard,” *Sleep Breath.*, vol. 23, no. 2, pp. 719–728, Jun. 2019, doi: 10/gqnkr4.
[77] U. J. Magalang *et al.*, “Agreement in the Scoring of Respiratory Events and Sleep Among International Sleep Centers,” *Sleep*, vol. 36, no. 4, pp. 591–596, Apr. 2013, doi: 10/gbct3c.
[78] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal Loss for Dense Object Detection,” p. 9, 2017, doi: 10/gf226d.

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

6. Supplementary Information

6.1. Supplementary Methods

Below is the additional description of how we developed the network and a more verbose description of its internals for deep learning specialists. However, all of the necessary methodological details of our evaluations are in the main Methods.

6.1.1. Hyperparameter search

Not only was the final network organically grown and modified during the hyperparameter search (e.g., adding layers, changing the number of features, etc.), but we also changed the inputs to the network. Before we settled on the basic structure of using a fully feed-forward network, we had also tried traditional machine learning techniques (support vector machines, naïve Bayes, etc.) and, later, various recurrent neural networks.

Those earlier attempts used many hand-crafted inputs that we assumed would be necessary based on our extensive signals processing experience and a review of similar literature. For instance, we had previously included various frequency band power summations to capture information linked to autonomic activity (e.g., lower frequencies for sympathetic activity and higher frequencies for parasympathetic activity). However, for the “traditional techniques”, these inputs were rarely informative for the models. Nonetheless, the predecessor to our current network initially included the full FFT spectrum of the ECG for each epoch and the autocorrelation of seven, 4-second windows for each epoch (to approximate HRV). However, we later found that the performance improved slightly if we removed the spectrum and autocorrelation inputs—leaving ECG as the only biophysical input.

Finally, before we settled on the feed-forward network, using the temporal convolution layers, we extensively tested two recurrent formulations: long short-term memory (LSTM) and gated recurrent unit (GRU). Additionally, we tried unidirectional (i.e., causal) and bi-directional variations. However, our results mirrored those of many later studies. Despite the adjective “long” in LSTM, these networks generally could not use contextual information from hundreds, let alone thousands, of epochs. Fortunately, the temporal convolution layer formulation also solves the issue of every recording being of a different length. Moreover, since our evaluations showed it was making better use of context, we decided to build future networks around that structure.

6.1.2. Additional meta-analysis details

As discussed, the starting point for the human-to-human PSG meta-analysis of kappa was a recently published meta-analysis study [7] (Methods 2.8). However, we had to exclude a few studies used in the previous analysis for various reasons. The first reason was that some studies only provided a single kappa value or a single contingency table (from which one could calculate a single kappa). In either case, the papers provided no information about the variance in the kappa values for multiple scorers; therefore, we could not derive a SE, SD, or 95% CIs. Four studies fell into this group: [27], [28], [29], [30]. Two of the studies had results that were duplicative of other results in the same study. First, in [69], the “all” group included “disorder” and “healthy”. Second, in [70], the “1EEG” result was from the same set of recordings as the “3EEG” result, just with two fewer electrodes. We tabulated the source data, as taken (directly or converted) from the source studies, in Supplementary Table S1 and Supplementary Table S2.

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

Supplementary Table S1. Studies for meta-analysis of overall kappa

Study	Group	Source	n	kappa	SE	SD	95% CI	Note
Norman, 2000 [69]	disorder	Fig. 2	10	0.606		0.057		1
	healthy	Fig. 2	10	0.648		0.117		1
Penzel, 2003 [71]		Table 1	8	0.470		0.150		1
Danker, 2004 [72]	GAD	Fig. 1	7	0.804		0.072		2
	GAD+INS	Fig. 1	11	0.770		0.102		2
	DEP	Fig. 1	9	0.683		0.126		2
	PLMS	Fig. 1	5	0.695		0.141		2
	OSAS	Fig. 1	51	0.679		0.106		2
	PD	Fig. 1	15	0.610		0.179		2
Danker, 2009 [55]	healthy	Fig. 5	56	0.730		0.095		2
	patient	Fig. 5	16	0.729		0.132		2
Ruehland, 2011 [70]	3EEG	Table 4	10	0.670	0.040			
Shambroom, 2012 [73]		Fig. 4	26	0.752		0.080		2
Elliot, 2013 [74]	pair 1/2	Table 4	16	0.572			0.550-0.582	3
	pair 2/3	Table 4	16	0.508			0.498-0.518	3
Zhang, 2015 [75]	control	Table 1	7	0.570		0.090		
	narcolepsy	Table 1	15	0.540		0.100		
	SAHS	Table 1	8	0.580		0.120		
Deng, 2019 [76]		Fig. 2	40	0.762		0.092		2

The table contains the input data we used for the meta-analysis. We collated it by study and included the data extracted, where the source data came from in the paper, and any additional notes. Note 1: We calculated the mean and SD from the boxplot values provided. Note 2: Quartile data provided by boxplots were converted to mean [32] and SD [33]. Note 3: Overall kappa values for Elliot, 2013 [74] were not in the original paper but were provided by the corresponding author.

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

Supplementary Table S2. Studies for meta-analysis of stage-specific kappas

Study	Group	Source	n	stage	kappa	SE	SD	95% CI	Note
Kunz, 2000 [43]		Table 1	172	Wake	0.821			0.708-0.886	
				N1	0.376			0.260-0.483	
				N2	0.759			0.694-0.810	
				N3	0.731			0.556-0.834	
				REM	0.874			0.809-0.918	
Danker, 2004 [72]		Fig. 2	196	Wake	0.786		0.149		2
				N1	0.359		0.171		2
				N2	0.692		0.137		2
				N3	0.660		0.249		2
				REM	0.846		0.097		2
Danker, 2009 [55]		Fig. 7	72	Wake	0.809		0.115		2
				N1	0.415		0.174		2
				N2	0.719		0.115		2
				N3	0.675		0.227		2
				REM	0.879		0.072		2
Ruehland, 2011 [70]	3EEG	Table 4	10	Wake	0.800	0.040			
				N1	0.400	0.030			
				N2	0.610	0.040			
				N3	0.600	0.060			
				REM	0.880	0.020			
Elliot, 2013 [74]	pair 1/2	Table 4	16	Wake	0.680			0.650-0.690	
				N1	0.120			0.100-0.130	
				N2	0.580			0.460-0.720	
				N3	0.760			0.700-0.820	
				REM	0.440			0.390-0.490	
	pair 2/3	Table 4	16	Wake	0.580			0.550-0.590	
				N1	0.080			0.060-0.100	
				N2	0.550			0.540-0.560	
				N3	0.200			0.140-0.230	
				REM	0.410			0.360-0.440	

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

Supplementary Table S2 (continued).

Study	Group	Source	n	stage	kappa	SE	SD	95% CI	Note
Magalang, 2013 [77]		Table 3	15	Wake	0.780			0.770-0.790	
				N1	0.310			0.300-0.320	
				N2	0.600			0.590-0.610	
				N3	0.670			0.650-0.690	
				REM	0.780			0.770-0.790	
Zhang, 2015 [75]	control	Table 1	7	Wake	0.650		0.120		
				N1	0.160		0.120		
				N2	0.580		0.110		
				N3	0.490		0.240		
				REM	0.790		0.180		
	narcolepsy	Table 1	15	Wake	0.580		0.150		
				N1	0.300		0.170		
				N2	0.550		0.130		
				N3	0.680		0.190		
				REM	0.660		0.140		
	SAHS	Table 1	8	Wake	0.760		0.120		
				N1	0.190		0.120		
				N2	0.500		0.190		
				N3	0.640		0.140		
				REM	0.690		0.160		
Deng, 2019 [76]		Fig. 2	40	Wake	0.888		0.061		2
				N1	0.456		0.141		2
				N2	0.725		0.111		2
				N3	0.776		0.138		2
				REM	0.871		0.067		2

The table contains the input data which we used for the meta-analysis. We collated it by study and included the data extracted, where the source data came from in the paper, and any additional notes. Note 2: Quartile data provided by boxplots were converted to mean [32] and SD [33].

6.2. Supplementary Results

In the following section, we present a collection of additional results that we believe bolster our central claims. Our approach has been to lean towards exhaustiveness, ensuring that others can thoroughly examine our main findings from all angles. This methodological choice reflects our commitment to transparency and our desire to facilitate a comprehensive interrogation of our results.

6.2.1. Meta-analysis tabulations

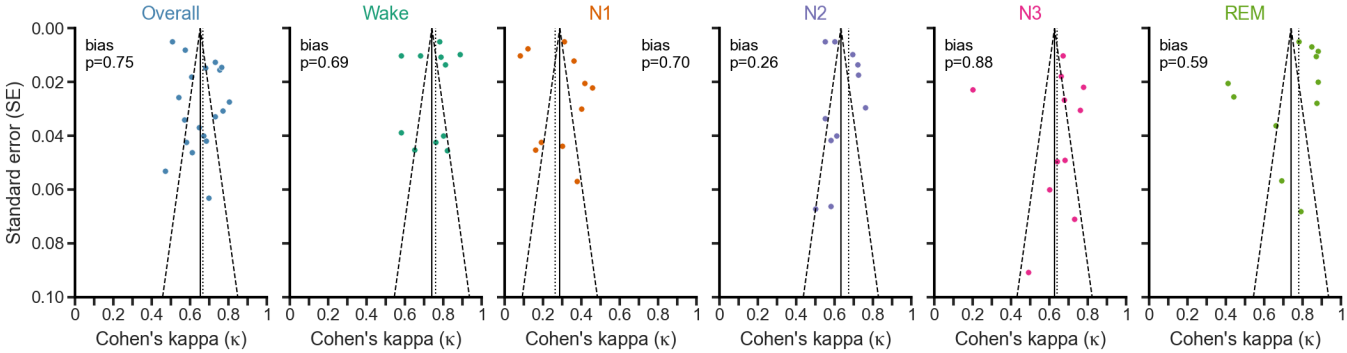
We tabulated the results of all of our random-effects meta-analyses in Supplementary Table S3. The large I^2 values suggest significant heterogeneity in the input data. We also tested for publication bias using visual and numerical techniques [36]. We found no bias in the meta-analysis inputs (Supplementary Fig. S1).

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

Supplementary Table S3. Meta-analysis results

	Overall	Wake	N1	N2	N3	REM
Study count (n)	19	11	11	11	11	11
Mean kappa (κ)	0.652	0.740	0.287	0.633	0.627	0.740
Lower 95% CI (κ)	0.597	0.672	0.196	0.581	0.510	0.664
Upper 95% CI (κ)	0.708	0.809	0.379	0.685	0.744	0.815
Lower 95% PI (κ)	0.384	0.487	-0.073	0.424	0.197	0.438
Upper 95% PI (κ)	0.928	*1.000	0.666	0.845	*1.000	*1.000
I^2	97.8	98.4	99.0	97.1	97.7	98.7
τ^2	0.012	0.010	0.017	0.005	0.028	0.012
Publication bias p-value	0.751	0.687	0.703	0.257	0.882	0.586

Tabulations of the meta-analysis kappas for Overall and each stage. *For three of the upper PIs, the kappa result was greater than 1.0. We instead report them as the maximum value kappa can obtain—1.0.



Supplementary Fig. S1. Funnel plots of meta-analysis inputs

We show the funnel plots of every input for each stage's random-effects analysis. For each stage, the solid vertical line is the random-effects estimate. The dashed diagonal lines are the 95% CIs of the estimate. We show the publication bias by the dotted vertical line (intercept point estimate) and the p-value.

6.2.2. Comparison with human-scored PSG

The adjusted p-values, using the Hochberg procedure, are in Supplementary Table S4.

Supplementary Table S4. Adjusted p-values for all t-tests

	Overall	Wake	N1	N2	N3	REM
Primary model (five datasets)	1.71E-53	3.72E-93	2.06E-67	1.29E-28	**1.000**	1.05E-36
CCSHS	1.51E-29	7.21E-45	6.34E-11	4.24E-15	1.58E-25	5.37E-23
CFS	1.83E-11	1.53E-21	3.42E-03	3.59E-04	4.77E-03	6.60E-03
CHAT	4.96E-46	1.47E-64	2.80E-57	1.36E-17	1.34E-58	1.66E-31
MESA	1.86E-06	9.22E-20	2.50E-21	5.17E-03	**1.000**	4.96E-04
WSC	**0.209**	2.77E-03	6.75E-07	1.08E-03	**1.000**	1.65E-08
Real-time model (five datasets)	5.77E-34	3.90E-71	9.10E-50	5.75E-14	**1.000**	1.40E-12

We show adjusted p-values for each non-inferior t-test. Using the Hochberg procedure, we adjusted the p-values to correct for multiple comparisons (for each sleep stage). We highlighted values with n.s. results in bold. All significant values are in scientific notation.

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

6.2.3. Comparison with EEG-less models

We tabulated all the EEG-less results we compared with and the boot-strapped samples we generated in Supplementary Table S5. We also show these same results in the main Results (Fig. 7).

Supplementary Table S5. Comparison with other EEG-less models

	Overall Cohen's kappa (κ) of all epochs (stage granularity)					
Model/Device (inputs)	**5-stage**	**4-stage**	**3-stage**	**Fig. marker**	**Reference**	**Year**
Ours (ECG)	0.726 [0.715, 0.736]	0.769 [0.759, 0.780]	0.842 [0.829, 0.848]			2024
(PPG)		0.650		1	Radha [46]	2021
(PPG, acti)		0.620 ±0.12	0.680 ±0.11	2	Wulterkens [47]	2021
(ECG, acti)		0.600		3	Fonseca [48]	2020
(HR)		0.660		4	Sridhar [49]	2020
(ECG, resp)	0.585		0.697	5	Sun [45]	2020
Fitbit (PPG, acti)		0.520		6	Beattie [50]	2017
(HRV)			0.610	7	Yoon [51]	2017
(ECG, acti)			0.580	8	Domingues [52]	2014
(HRV, acti, resp)		0.560	0.620	9	Willemen [53]	2014
(PPG, SpO_2, resp)	0.510 ±0.01			10	Sady [54]	2013

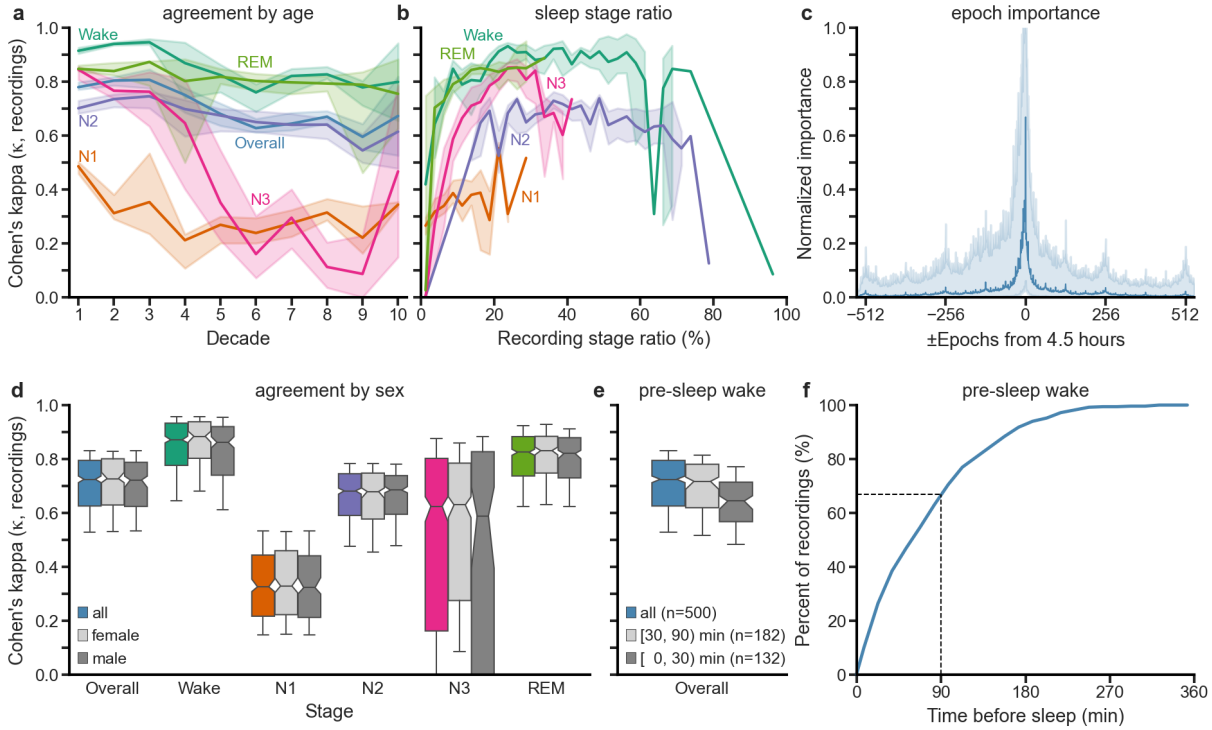
Our model's median and 95% CIs of each stage granularity are from the bootstrapped sampled. The other models perform significantly worse than the current. Models are in chronological order. resp=respiration. acti=actigraphy. five-stage: W/N1/N2/N3/REM, four-stage: W/"Light"/"Deep"/REM, three-stage: W/NREM/REM. Most sources did not provide SD or CI values.

6.2.4. Additional investigations on testing set

We performed other assessments of the model's performance on the original testing set. We did these investigations to understand better the main results (e.g., to determine why any differences exist). Doing so could help us understand under what conditions the network is best suited and when and why it might perform worse than expected.

The initial series of results were on the original testing set. First, the model's performance is largely unaffected by age except for N3, which decreases beginning in the fifth decade (Supplementary Fig. S2a). Our results match the expected consequences for stage ratios that are exceptionally low or high (Supplementary Fig. S2b, [23]). Similar to the finding on stage ratios, the amount of pre-sleep wake (i.e., time awake between when the recording started and the subject fell asleep) was important. Salience computations showed that the network uses contextual information before and after to score that epoch (Supplementary Fig. S2c). Finally, the subject's sex did not affect the performance (Supplementary Fig. S2d). We found that the performance deteriorated when there were fewer than 30 minutes of pre-sleep wake (Supplementary Fig. S2e, the distribution of pre-sleep wake is in Supplementary Fig. S2f).

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545



Supplementary Fig. S2. Performance on the testing set

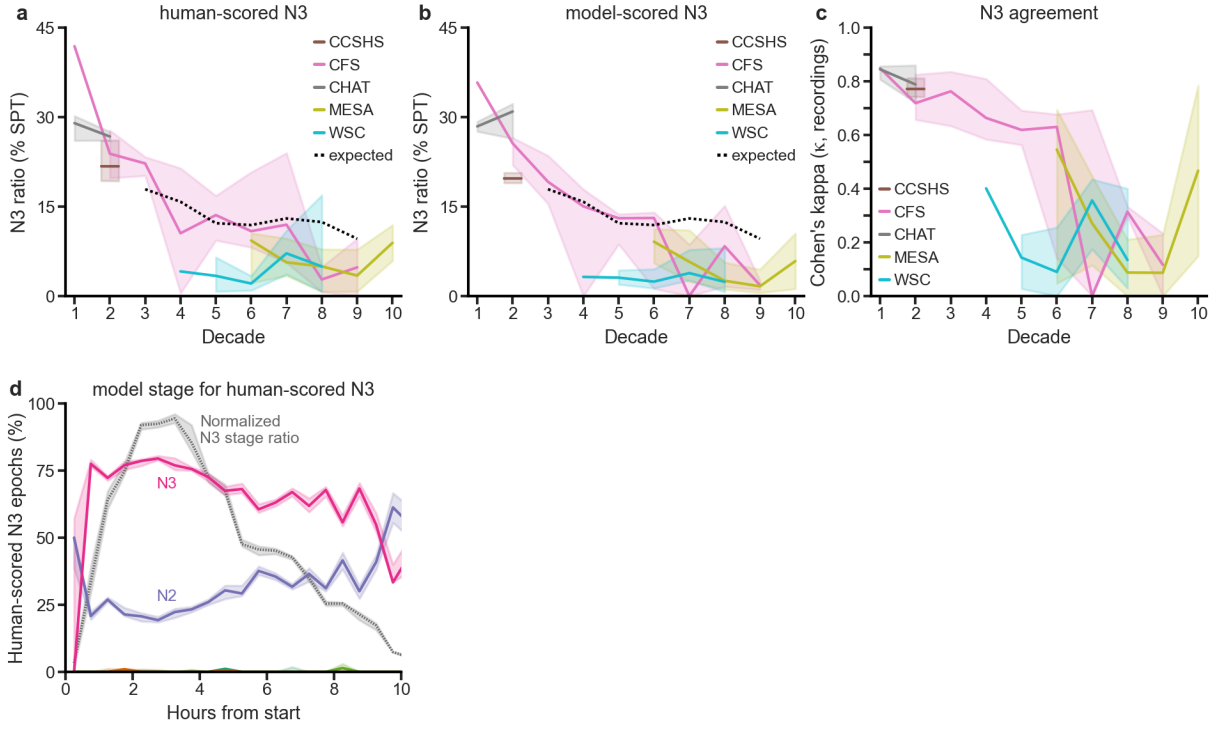
(a) When stratified by age (decade 1=age 0-9yr.), REM kappa is the least affected by age, while N1 and N3 are the most affected, suggesting the broad applicability of the model. (b) Stratified by the recording's stage ratio, namely the proportion of recorded time spent in each stage, a stage's kappa decreased when its ratio was minuscule—or nearly everything. (c) Using integrated gradients (Methods 2.12) to determine the relative importance of the epochs used to score at epoch 540 (4.5 hours from the start). This result reveals that the baseline network (Fig. 2) uses past and future epochs, with noticeable "blips" at powers of two distance from the current epoch. The line represents the median importance of n=500 recordings, and shaded areas represent the 95% CIs. (d) When disaggregating by sex, results are essentially the same for each stage, with noticeable but insignificant sex differences for Wake and N3. (e) Stratified by the amount of pre-sleep wake, kappa is significantly lower for recordings with less than 30 minutes of pre-sleep wake. Prospective users of the model should plan accordingly. Most recordings (n=314) contain less than 90 minutes. (f) A majority (n=314) of recordings contain less than 90 minutes of pre-sleep wake. Shaded areas represent 95% CIs. Whiskers at P10 and P90.

6.2.5. Deeper investigations of N3

Our primary comparison with human-scored PSG (Fig. 4 and Fig. 5) shows that the N3 performance was the single stage that could not achieve statistical significance. However, we disaggregated the results by source dataset to show that this non-significant result was primarily a result of the MESA and WSC datasets. Therefore, we decided to investigate this further to see if there were any hints about what was different about these datasets. To this end, we show that the human-scored N3 stage ratios for MESA and WSC are well below the expectation for their subjects' ages (Supplementary Fig. S3a). Moreover, we saw this same discrepancy for the model-scored N3 stage ratios for MESA and WSC (Supplementary Fig. S3b). As mentioned, when the proportion of a stage tends toward zero, the stage-wise kappa will also quickly decrease toward zero. We also found the same when stratified by decade and disaggregated by dataset (Supplementary Fig. S3c). We discuss these findings at length in Discussion.

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

Additionally, we found an interesting trend when looking at the performance of our primary model across the recording (Supplementary Fig. S5a). The N3 performance started to decrease quickly around 9 hours after the start of recording. We later found that when looking at the human-scored N3 across the night, the model agrees about 75% of the time around 3 hours in the recording. However, this agreement gradually decreases until around 9 hours. Eventually, the model scores more of the epochs as N2—that the human had scored as N3 (Supplementary Fig. S3d). This result could indicate some fundamental difference in N3 that occurs later in the night, or it could be an artifact of the lower-than-expected N3 stage ratio.



Supplementary Fig. S3. Investigations into N3 scoring

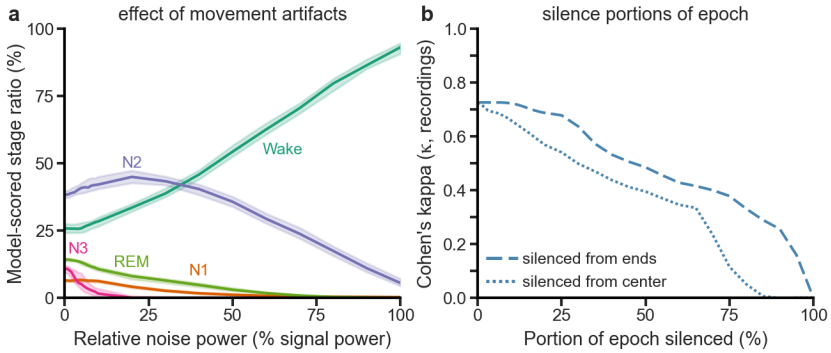
(a) The human-scored N3 stage ratios for the testing set, disaggregated by source dataset and decade (similar to Fig. 1e, except only the testing set recordings). The dotted black line is the expected median N3 ratio from an n=198 study [57]. (b) The model-scored N3 stage ratios for the testing set are in the same format as panel a. (c) The N3 agreement between humans and models, disaggregated by source dataset. The worst performing decade/study combinations occur where the proportion of N3 predicted by either the human or model is significantly lower than the expected N3 ratio. (d) Looking only at human-scored N3 epochs, the model predominantly scores them as N3 until around 9 hours in and afterward as N2. The gray line is the normalized likelihood of N3 epochs (i.e., where in the night the human scored epochs as N3, Fig. 1d). Shaded areas represent the 95% CIs.

6.2.6. Additional robustness investigations

We conducted two additional robustness analyses. As mentioned in the main Results, substantial noise levels are necessary to affect the performance meaningfully (Fig. 10b). The exception is movement artifact noise, which causes a steeper roll-off. When looking at how the model scores those epochs corrupted by movement artifacts, we see that it increasingly scores epochs as Wake (Supplementary Fig. S4a).

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

Secondly, similar to the results presented earlier on silencing whole epochs (Fig. 10e), we silenced portions of the input for each epoch to determine what portions of an individual epoch the network uses in constructing its features. The results show that the network is more affected by losing the data from the epoch's center (dotted line, Supplementary Fig. S4b) than from its ends (dashed line). Moreover, the network can lose about 5% from each end (10% total) before the performance begins to degrade.



Supplementary Fig. S4. Additional robust performance findings

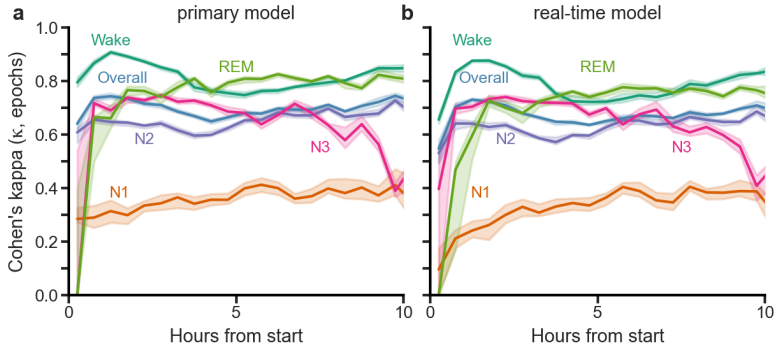
(a) When adding movement artifact noise (Fig. 10b), with increasing noise power, the network scores more epochs as Wake. (b) When silencing increasing portions of each epoch, starting from both ends of each epoch, more silence leads to worse performance. Performance is worse when silencing from the center than from the ends. Shaded areas represent the 95% CIs.

6.2.7. Additional real-time investigations

Since the real-time model would score in real-time, it would be helpful to know how it performs across time. Additionally, we performed the same analysis on our primary model to provide some baseline for that result. In other words, we wanted to assess the performance of scoring an individual epoch as a function of its location in the recording. We found that, although the likelihood of a particular stage varies across the night, the likelihood had a negligible effect on performance. The exception was the classification of N3 and REM right at the beginning and N3 after about 9 hours (Supplementary Fig. S5a).

Next, when performing this same analysis on the real-time model, we see a slight uniform decrease and some changes at the beginning of the night (Supplementary Fig. S5b, the stage markers are in the same locations across panels to make comparisons easier). It takes slightly longer for N1 and N3 performance to rise to nearly the same levels as they were for our primary model.

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545



Supplementary Fig. S5. Performance across time for both models

(a) When stratified by time (30-minute windows across all recordings), although there is some variation in kappas, they are relatively consistent. Furthermore, the stage-specific kappas fall in a narrow range—matching the average performance (Fig. 4 and Fig. 5). However, the lower model performance is because of a lack of diverse observations in those periods (e.g., REM is rare at the beginning of the night, whereas N3 becomes less frequent after eight hours Fig. 1d). (b) Using the same calculation as panel a, but from the results of the real-time model. We can see a slight but fairly uniform decrease in performance for all stages. Of note is that the performance of N1 and REM takes slightly longer to "ramp up" versus our primary model. The stage markers are in the same locations as panel a to make comparisons easier. Shaded areas represent the 95% CIs.

6.2.8. Summary tabulations

This subsection provides detailed tabulations to present our analytical findings comprehensively. Table Supplementary Table S6 outlines the epoch counts for each dataset set, offering a clear view of the iterations undertaken during our analysis. Following this, Supplementary Table S7 presents Cohen's kappas for the testing set, detailing the agreement levels between raters and the consistency of our classification methods.

Lastly, Supplementary Table S8 has a tabulation of additional classification metrics. However, we remind the reader that Cohen's kappa is the only appropriate and commonly reported metric. Moreover, while it is mathematically possible to use the contingency tables to compute the values of other classification metrics, these other metrics assume the existence of a ground truth. This assumption is not the case for sleep stages. See the Cohen's kappa section in Methods 2.6.

Supplementary Table S6. Epoch counts for each set

	Epoch counts for each stage						
Set	**Wake**	**N1**	**N2**	**N3**	**REM**	**Unscored**	**Total**
Training	980,338	209,550	1,315,130	457,009	447,811	41,387	3,451,225
Validation	165,513	34,830	218,264	74,034	71,904	7,223	571,768
Testing	165,300	34,601	220,642	73,566	73,032	7,209	574,350
Total	1,311,151	278,981	1,754,036	604,609	592,747	55,819	4,597,343

For the 4,000 recordings selected, above are the epoch counts for each stage and unscored epochs.

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

Supplementary Table S7. Cohen's kappas on testing set

	Cohen's kappa (κ)					
Kappa calculation method	**Overall**	**Wake**	**N1**	**N2**	**N3**	**REM**
Median kappa of all recordings (n=500)	0.725	0.871	0.326	0.682	0.625	0.825
Mean kappa of all recordings (n=500)	0.697	0.830	0.333	0.651	0.505	0.777
Kappa of all epochs (n=567,141)	0.726	0.862	0.373	0.671	0.703	0.805

There are slight differences in final values depending on the method used to aggregate and calculate the results. We report all three here to enable direct comparisons with a wider variety of literature.

Supplementary Table S8. Additional classification metrics

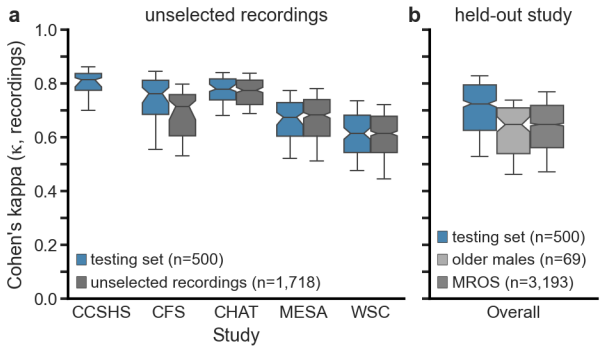
	Mean of value for all recordings					Stage-weighted
Measure	**Wake**	**N1**	**N2**	**N3**	**REM**	**average**
Accuracy	0.943	0.921	0.840	0.935	0.954	0.893
Recall	0.864	0.440	0.801	0.548	0.840	0.797
Precision	0.899	0.361	0.795	0.589	0.781	0.826
Specificity	0.967	0.952	0.860	0.968	0.969	0.920
F1-score	0.872	0.372	0.791	0.517	0.796	0.798

Cohen's kappa is the appropriate measure of inter-rater agreement (because it does not assume a "ground truth" and is the most commonly reported inter-rater statistic). While it is mathematically possible to calculate other classification measures from the same stage-wise contingency tables, we remind the reader that these measures assume the human-provided scores are correct. However, any two expert human scorers will score differently. Therefore, the assumption of any single scorer being "correct" is dubious. The values listed are the mean for all n=500 recordings. We computed the stage-weighted averages using the stage ratios for each recording individually.

6.2.9. Held-out results

We also evaluated the model on additional recordings to test if there were unique attributes about the recordings selected or if study-specific learning had occurred. Either issue would hamper generalizability. The first collection includes the recordings from the original studies that met the quality criteria (Methods 2.2.3) but that we did not randomly select (Supplementary Fig. S6a). It bears stressing that we did not add these recordings to our testing set because it would skew the age and sex distributions from the target distribution. These results show no significant difference between the recordings we had selected or omitted (i.e., unselected). The second collection comes from a study, MROS, which we did not use for the training, validation, or testing sets. Their performance was equivalent to the age-matched males (decades ≥ 7) in the testing set (Supplementary Fig. S6b). This result indicates that if the model had learned any study-specific features, these features were unnecessary for adequate performance.

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545



Supplementary Fig. S6. Unselected recordings and held-out study

(a) We did not entirely use four of the five studies from which we sampled recordings (Methods 2.14). The results of those unselected recordings show no differences from those of their counterparts in the testing set. CFS (n=105), CHAT (n=270), MESA (n=498), WSC (n=845). (b) One study, MROS (n=3,193), which we did not use during the training phase, so we could evaluate it to determine if study-specific learning had occurred (i.e., learning features specific to the study apparatus or pipeline). The performance aligns with the aged-matched males that constituted the study's demographics (decades ≥ 7) from the testing set (light blue, from CFS, MESA, and WSC). Whiskers at P10 and P90.

6.2.10. Loss function evaluations

Although we did not comprehensively evaluate our loss function, we did evaluate it against the most commonly used classification loss functions (Supplementary Table S9). We found that while the default cross-entropy function sometimes performed equal to or better than our loss function (max +1%), it did so at the expense of N1 (-27%).

Furthermore, it was worth evaluating the assumption that a single model that classifies all five stages at the same time could perform better than a model of the same size that only has to classify a single stage (e.g., Wake vs. not-Wake). We found that for each of the five stages, a single five-stage model performs as well or better than a collection of one-stage models (Supplementary Table S10).

Supplementary Table S9. Loss function comparisons

	Cohen's kappa (κ) of all epochs					
Loss function	**Overall**	**Wake**	**N1**	**N2**	**N3**	**REM**
Class kappa mean (ours)	0.726	0.862	0.373	0.671	0.703	0.805
Cross-entropy	0.734	0.867	0.274	0.682	0.699	0.805
Cross-entropy (weighted)	0.669	0.845	0.332	0.583	0.677	0.786
Focal [78]	0.732	0.862	0.297	0.679	0.703	0.801
Cohen's kappa (overall)	0.720	0.854	0.000	0.669	0.697	0.795
Ratio of ours to best	99%	99%	100%	98%	100%	100%

Although some loss functions achieve slightly better Overall performance, N1 performance is markedly worse. We highlighted the highest performance for each column in gray. The weights provided for weighted cross-entropy were the inverse of the stage proportions in the n=3,000 training set (i.e., [0.214, 1.000, 0.159, 0.459, 0.468]).

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

Supplementary Table S10. Five-stage versus one-stage comparisons

	Cohen's kappa (κ) of all epochs				
Loss function	**Wake**	**N1**	**N2**	**N3**	**REM**
Complete 5-stage (default)	0.862	0.373	0.671	0.703	0.805
Best 1-stage (each)	0.862	0.353	0.671	0.692	0.800
Ratio of 5-stage to 1-stage	100%	106%	100%	102%	101%

When the loss function classifies all five stages together (i.e., the geometric mean), it performs equal to or better than any stage-specific kappa that one might optimize individually.

6.3. Supplementary Discussion

In the subsequent section, we offer additional analysis that, though not directly central to our main assertions, bears tangential relevance to our study's overarching themes. This supplementary discussion provides insights into the loss function comparison, the issues we discovered with numerous EEG-less studies, and possible future directions.

6.3.1. Loss function comparison

Since we developed our loss function during the hyperparameter search, the contemporaneous results of each loss function would make little sense when compared with the results presented here. Therefore, we re-ran the training using the final model with three common loss functions for classification, as well as the overall Cohen's kappa. It is important to remember that it is probable that the loss function itself influenced the network's evolution. Therefore, the final network might be less performant while training with any other loss function. However, the reported results mirror the relative values we saw during development. On balance, during the hyperparameter search, we examined several dozen ways of combining and weighting various loss functions to little avail. Most loss functions either ignored N1 (i.e., $\kappa = 0$) or could not bring N1's performance up to what we found was achievable with our loss function.

The results in Supplementary Table S9 show that unweighted cross-entropy and focal loss have a slightly higher overall kappa (+1%, both). However, their N1 performance is significantly worse (-62 and 59%, respectively). Given that our loss function had significantly better N1 performance versus the marginal decrease in overall performance, we think it is a worthwhile tradeoff.

6.3.2. Exclusion of some EEG-less studies

One of the issues with some EEG-less studies mentioned in the main Discussion bears more explanation. This issue is the contamination of the evaluation set. The issue is a serious methodological problem that comes in two forms and is more common than expected. The first form of this issue is using so-called "subject-specific" classifiers. The researchers trained and evaluated these models on data from the same recordings, whereby they assigned an individual epoch to either the training or evaluation set. The problem is that the data will be nearly identical between adjacent or nearby epochs. Therefore, the evaluation data is highly similar to the training data. The second form of the issue is using a single evaluation set. There should always be two evaluation sets, a validation set, and a hold-out testing set. Crucially, researchers should not evaluate the model on the hold-out testing set until—after—they have selected a final model. During the development of any model, hyperparameter tuning is necessary to achieve the best-performing model. To improve the model and converge on the best hyperparameters, researchers use a validation set that is different from the training set. However, researchers also sometimes

Published in Computers in Biology and Medicine, doi: 10.1016/j.compbiomed.2024.108545

conduct this step using cross-validation. The problem is that the hyperparameter tuning process "leaks" information from the validation set into the model (i.e., the researchers make model choices based on the performance of the validation set). Since the goal is generalizability, testing must estimate the model's performance on unseen data. Unfortunately, if researchers use the same validation set (or the same cross-validation population) for testing, they are unwittingly evaluating the performance on already-seen data. Reviewing the literature requires carefully reading the methods to notice these issues. It is often only obvious when the results mention an "external" or "unseen" data evaluation, where the performance is often significantly worse than their top-line numbers.

6.3.3. Future directions

The ability of our model to score sleep stages using a single lead of ECG on par with experienced human scorers using PSG data raises several questions. The most salient question is what specifically in the input data is the network using to such a pronounced effect. As mentioned, other EEG-less models have been mining downstream measures of ECG, such as HRV, with limited success compared to PSG performance. Moreover, in an earlier iteration of the network, we used additional inputs, including a surrogate for HRV—with no improvement in performance. We would like to investigate what the network is using.

Finally, it is worth highlighting that we only took one preemptive measure to improve the model's robustness while training. While spot-checking the input ECG data, we noticed from the waveform appearance that sometimes the technician had connected the electrodes backward (i.e., incorrect polarity). Instead of manually verifying all recordings, we had the data loader invert the ECG during training with a 50% probability (Methods 2.5). This operation undoubtedly made learning more difficult, forcing the network to develop a polarity-insensitive feature extraction. We could use this technique to improve the model's robustness in future iterations. Specifically, we could add noise (e.g., Gaussian), remove portions of epochs, or even remove entire epochs. We emphasize that there will likely be tradeoffs between incorporating these measures and the training time and final performance.