RESEARCH ARTICLE

Detection of sleep apnea from single-channel electroencephalogram (EEG) using an explainable convolutional neural network (CNN)

Lachlan D. Barnes[1☯], **Kevin Lee**[1☯], **Andreas W. Kempa-Liehr**[2], **Luke E. Hallum**[1]*

1 Department of Mechanical and Mechatronics Engineering, University of Auckland, Auckland, New Zealand, **2** Department of Engineering Science, University of Auckland, Auckland, New Zealand

☯ These authors contributed equally to this work.
* l.hallum@auckland.ac.nz

Abstract

Sleep apnea (SA) is a common disorder involving the cessation of breathing during sleep. It can cause daytime hypersomnia, accidents, and, if allowed to progress, serious, chronic conditions. Continuous positive airway pressure is an effective SA treatment. However, long waitlists impede timely diagnosis; overnight sleep studies involve trained technicians scoring a polysomnograph, which comprises multiple physiological signals including multi-channel electroencephalography (EEG). Therefore, it is important to develop simplified and automated approaches to detect SA. In the present study, we have developed an explainable convolutional neural network (CNN) to detect SA events from single-channel EEG recordings which generalizes across subjects. The network architecture consisted of three convolutional layers. We tuned hyperparameters using the Hyperband algorithm, optimized parameters using Adam, and quantified network performance with subjectwise 10-fold cross-validation. Our CNN performed with an accuracy of 69.9%, and a Matthews correlation coefficient (MCC) of 0.38. To explain the mechanisms of our trained network, we used critical-band masking (CBM): after training, we added bandlimited noise to test recordings; we parametrically varied the noise band center frequency and noise intensity, quantifying the deleterious effect on performance. We reconciled the effects of CBM with lesioning, wherein we zeroed the trained network's 1st-layer filter kernels in turn, quantifying the deleterious effect on performance. These analyses indicated that the network learned frequency-band information consistent with known SA biomarkers, specifically, delta and beta band activity. Our results indicate single-channel EEG may have clinical potential for SA diagnosis.

Citation: Barnes LD, Lee K, Kempa-Liehr AW, Hallum LE (2022) Detection of sleep apnea from single-channel electroencephalogram (EEG) using an explainable convolutional neural network (CNN). PLoS ONE 17(9): e0272167. https://doi.org/10.1371/journal.pone.0272167

Editor: Humaira Nisar, Universiti Tunku Abdul Rahman, MALAYSIA

Received: November 29, 2021

Accepted: June 21, 2022

Published: September 13, 2022

Data Availability Statement: We used three datasets. Our convolutional neural network (CNN) was trained and tested using data drawn from Sleep Health Heart Study (SHHS) Visit 2 [19, 20]. This dataset is publicly available via the National Sleep Research Resource which provides access to large collections of de-identified physiological signals for research purposes (https://sleepdata.org) [19]. We performed further testing of our trained CNN using data from the St. Vincent's University Hospital / University College Dublin Sleep Apnea Database [21], and the MIT-BIH

Introduction

Sleep apnea (SA) is a progressive disease which involves repeated episodes of apnea and/or hypopnea during sleep. It afflicts between 2% and 7% of the general population [1, 2] and

Polysomnographic Database [21, 22]. These latter two datasets are publicly available via PhysioNet which provides access to a collection of de-identified physiological signals for research purposes (https://www.physionet.org) [21]. The University of Auckland Human Participants Ethics Committee waived the requirement for any further approval to use these datasets in this study. 19. Zhang GQ, Cui L, Mueller R, Tao S, Kim M, Rueschman M, Mariani S, Mobley D, Redline S. The National Sleep Research Resource: towards a sleep data commons. J Am Med Inf Assoc. 2018 Oct 1;25(10):1351-1358. 20. Quan SF, Howard BV, Iber C, Kiley JP, Nieto FJ, O'Connor GT, Rapoport DM, Redline S, Robbins J, Samet JM, Wahl PW. The Sleep Heart Health Study: design, rationale, and methods. Sleep. 1997 Dec;20(12):1077-85. 21. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 2000 Jun 13;101(23):E215-20. 22. Ichimaru Y, Moody GB. Development of the polysomnographic database on CD-ROM. Psychiatry Clin Neurosci. 1999 Apr;53(2):175-7. Digital object identifiers for datasets are as follows: 1. Sleep Health Heart Study (SHHS) Visit 2: https://doi.org/10.25822/ghy8-ks59 2. St. Vincent's University Hospital / University College Dublin Sleep Apnea Database: https://doi.org/10.13026/C26C7D 3. MIT-BIH Polysomnographic Database: https://doi.org/10.13026/C23K5S.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

causes fragmented sleep. Sufferers often experience daytime hypersomnia, cognitive dysfunction that accompanies sleepiness, and an increased risk of workplace and motor vehicle accidents [3]. If allowed to progress, SA is associated with a range of serious, chronic conditions, including cardiovascular and cerebrovascular disease, and diabetes [1, 2, 4, 5]. Continuous positive airway pressure (CPAP)—the gold standard SA treatment—is highly effective [4]. However, there are barriers to SA detection, diagnosis, and treatment. The primary clinical tool for SA detection and diagnosis is overnight polysomnography (PSG), which involves sleep studies and manual scoring of recorded physiological signals by trained healthcare professionals [6]. PSG is instrumentation-intensive—it typically involves monitoring nasal or oral airflow; thoracic and/or abdominal movement; snoring; oxygen saturation; multi-channel electroencephalogram (EEG); electrooculogram (EOG); electrocardiogram (ECG); and, electromyogram (EMG)—and therefore, may interfere with standard patterns of sleep [7]. Overnight PSG is typically followed by manually titrated CPAP therapy [1].

Demand for overnight PSG exceeds supply; wait times for those requiring screening and diagnosis can range from 2 to 60 months in selected developed countries [8], and it is estimated that a large proportion of SA sufferers remain undiagnosed [9]. Tools to aid detection, diagnosis, and treatment of SA are therefore a worthy pursuit, especially those involving simplified instrumentation and automation, with potential for use in the home as well as the clinic [10–12]. Automated systems making use of explainable machine learning [13] could, potentially, be used to expedite and augment diagnostic and treatment decisions; in general, explainable systems are those wherein mechanisms of detection and/or classification are made available to clinicians. Automated detection systems have been explored [14] using various modalities such as ECG [15, 16], oxygen saturation [17], or airflow [18–20]. EEG contains a rich variety of physiological signals enabling objective measurement of sleep stage, as well as various sleep disorders (e.g., narcolepsy [21]). Hence EEG-based detection may have far-reaching applications in simplified at-home sleep tests and may offer a constructive perspective to present sleep monitoring modalities. It is also noteworthy that substantial progress has been made to improve the instrumentation and applicability of EEG for wearable devices [12] and the production of consumer level products [22] in pursuit of better health monitoring and sleep assessment.

The traditional approach to SA detection from EEG involves computation of features (e.g., energy and energy variance [23]) within predefined frequency bands [23, 24]. Features are concatenated to form a high-dimensional feature vector for use in classification. Convolutional neural networks (CNNs) are a form of artificial neural network, loosely inspired by hierarchical, computational models of visual processing in the cerebral cortex (review by LeCun et al. [25]). CNNs use convolution as a form of shift-invariant feature extraction, and learn, through training, to extract salient features from time series signals (or images) that are useful for classification. Recently, CNNs have demonstrated proficiency for the classification of images [26] and signals across a range of domains, including multi-channel EEG [27]. In contrast to traditional approaches, the CNN we develop here requires no postulation of features and frequency bands at the outset, meaning that features not traditionally associated with SA could be learned during the training procedure. Beside CNNs, there are several other state-of-the-art machine learning algorithms such as residual neural networks [20], and transformer architectures [28]. Although such architectures may offer superior performance in some domains compared to CNNs, it can come with the cost of explainability; explainability is a requirement for healthcare applications [29]. Furthermore, our CNN uses single-channel EEG; we are interested in the feasibility of a wearable sleep assessment device, analogous to recent single-lead electrocardiographic adhesive patches for detecting cardiac arrhythmia [30].

We hypothesized, first, that there is information in single-channel EEG enabling reliable subjectwise detection of SA by a CNN; by "subjectwise", we mean a CNN trained using data collected from a cohort of subjects, 1 through N, should generalize to detect SA in a previously unseen subject, N+1. Second, we hypothesized that knowledge of sleep stage (e.g., rapid-eye-movement sleep) should improve this SA detection. This hypothesis is reasonable because both SA as well as sleep stage are accompanied by characteristic alterations of EEG, and SA is associated with sleep stage (a point we elaborate in the Discussion). Third, we hypothesized that the network features enabling SA detection should be consistent with known SA biomarkers. To test these hypotheses, we trained a CNN to detect SA using single-channel EEG. To explain the trained CNN's mechanisms, we used two visualization techniques—critical-band masking [31, 32] (wherein band-limited noise was added to signals used to test the trained network) and filter lesioning [33] (wherein 1st-layer filter kernels comprising the trained network were zeroed in turn, and the effect on performance was quantified).

Materials and methods

Datasets

We used three datasets. Our CNN was trained and tested using data drawn from Sleep Health Heart Study (SHHS) Visit 2, which contains overnight EEG recordings from 2,650 patients sampled at either 125 Hz or 128 Hz [34, 35]. This dataset is publicly available via the National Sleep Research Resource which provides access to large collections of de-identified physiological signals for research purposes (https://sleepdata.org) [34]. We performed further testing of our trained CNN using data from the St. Vincent's University Hospital / University College Dublin Sleep Apnea Database [36], consisting of recordings from 25 participants sampled at 128 Hz, and the MIT-BIH Polysomnographic Database [36, 37], consisting of usable recordings from 16 patients sampled at 250 Hz. These latter two datasets are publicly available via PhysioNet which provides access to a collection of de-identified physiological signals for research purposes (https://www.physionet.org) [36]. The University of Auckland Human Participants Ethics Committee waived the requirement for any further approval to use these datasets in this study.

Sleep apnea annotation procedures differed considerably between datasets. These differences presumably hindered our ability to develop a single CNN using the SHHS data set which generalised to both the UCD and MIT data sets (see Discussion). The SHHS dataset scored sleep apneas using internally developed rules [35], and the scoring procedure involved multiple stages of both manual and computer-automated annotation. Few details regarding the UCD and MIT scoring were available. Furthermore, MIT data were pre-segmented (30-second segments) and pre-labelled, whereas SHHS and UCD data were annotated at finer temporal resolution.

The same duration criterion (i.e., 10 seconds) was applied to all SHHS data irrespective of the apnea type; physiological criteria were used to distinguish the apnea types (Fig 1). Because the duration criterion for UCD was unknown, we applied the SHHS criterion. Therefore, for SHHS and UCD data sets, we first divided the data into 30-second segments and then labelled all segments that contained at least 10 continuous seconds of "obstructive sleep apnea", "central sleep apnea", "mixed apnea", or "hypopnea" as, simply, "apnea". All other segments were labelled "non-apnea". For the MIT data set, we used the labels provided. All three data sets scored sleep stages using the Rechtshaffen and Kales criteria [38], albeit the SHHS dataset used a modified version of these criteria [35]. Channel C4-A1 [39] was the only channel present across all three datasets and therefore was used exclusively to ensure consistency.

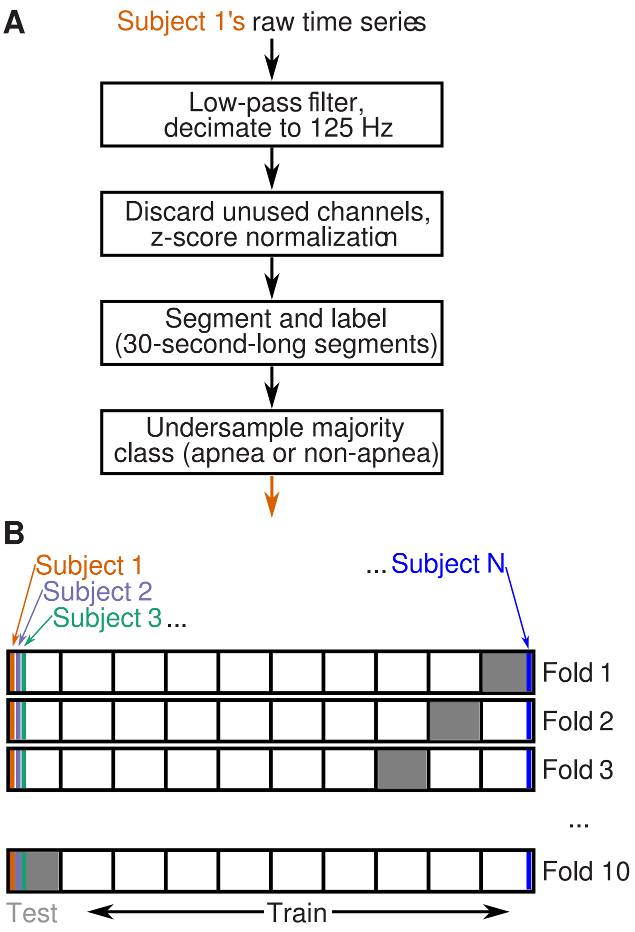
A
Subject 1's raw time series
Low-pass filter, decimate to 125 Hz
Discard unused channels, z-score normalization
Segment and label (30-second-long segments)
Undersample majority class (apnea or non-apnea)
B
Subject 1
Subject 2
Subject 3 ...
... Subject N
Fold 1
Fold 2
Fold 3
...
Fold 10
Test
Train

Fig 1. Data processing and subjectwise 10-fold cross-validation. (A) We used recordings from the SHHS dataset [34, 35]. For each subject, we low-pass filtered, downsampled, normalised, segmented, labelled, and undersampled recordings as indicated. (B) On each fold of our 10-fold cross-validation procedure, any subject's recordings appeared only in the training set (white) or the testing set (gray). For example, subjects 1, 2, and 3 contributed to the training sets of folds 1 to 9, but were excluded from the training set of fold 10, in which subjects 1, 2, and 3 were part of the test set. Therefore, overall, our assessment of the model's performance captured its ability to generalize across subjects.

https://doi.org/10.1371/journal.pone.0272167.g001

Classifier design

Schirrmeister and colleagues [27], evaluated shallow, deep, hybrid and residual convolutional neural networks, as well as the filter bank common spatial patterns (FBCSP) algorithm. These algorithms were evaluated across several multi-channel EEG motor-decoding datasets. They concluded that shallow and deep architectures reached, and sometimes exceeded the performance of FBCSP (the de facto standard), in contrast to hybrid and residual neural networks which fell short of the performance of FBCSP. In light of their findings, we constructed our CNN network (Fig 2) using three convolutional layers; each convolutional layer involved normalization, exponential linear unit (ELU) activation, and max-pooling. We kept our network architecture simple with a relatively low number of convolutional layers, as we anticipated that it would better facilitate explainability (i.e., mechanistic analysis of the trained network's performance, described below). We used batch normalisation (after each convolution layer) and dropout (after each pooling layer) [40] to improve network stability and safeguard against overfitting. As illustrated in Fig 2, the last convolutional layer was followed by a dense layer, and an output layer with the softmax activation function. We initialized the weights of our

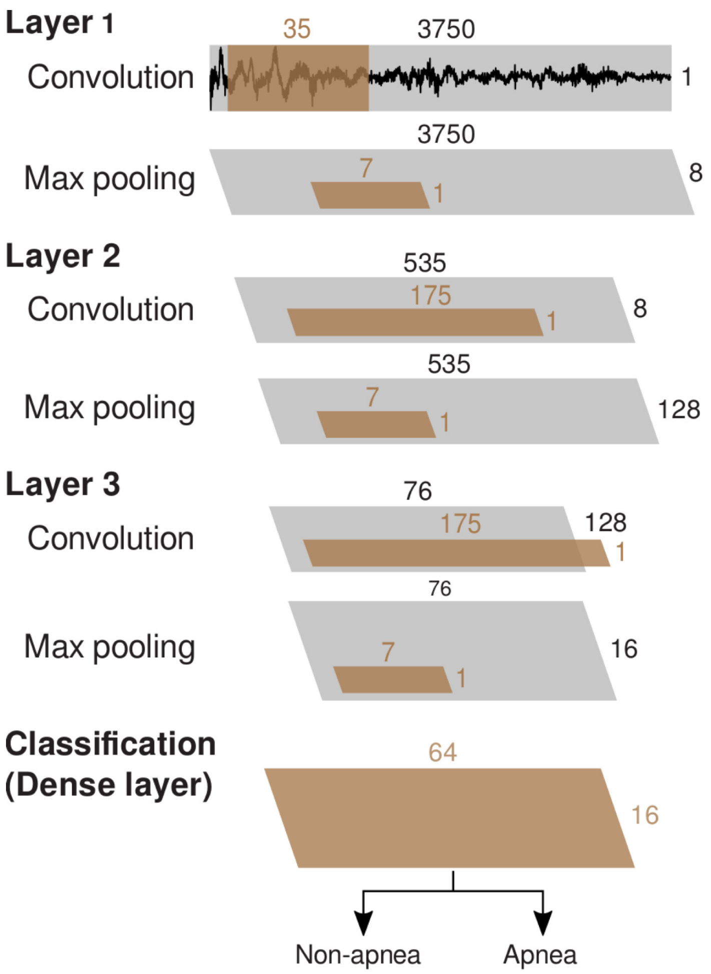

Fig 2. The architecture of our CNN trained to detect SA, comprising three convolutional layers. Convolutions had a stride size of one and used zero padding. Each convolutional layer was followed by batch normalisation, ELU activation, and dropout (these operations are not illustrated). The dense layer was preceded by a flattening operation, and followed by ELU activation and a dropout layer. The output layer used a softmax classifier [42]. (Symbology after Schirrmeister and colleagues [27]).

https://doi.org/10.1371/journal.pone.0272167.g002

CNN by drawing from a truncated normal distribution with zero mean. To train the CNN, we minimized the cross-entropy loss function through backpropagation. Backpropagation was optimized using Adam [41] in the standard fashion: alpha coefficient was set at 0.9, and beta coefficient at 0.999. The learning rate was tuned alongside other hyperparameters (see Table 1).

Hyperparameter tuning

There are few studies describing how best to optimize network hyperparameters (i.e. network variables set prior to training) to detect SA in EEG. Therefore we approached hyperparameter optimization in a semi-automated manner. First, we reviewed literature that reported hyperparameter values, and we constructed a hyperparameter search space encompassing the values we found. Second, we systematically searched this space for optimal hyperparameter combinations. We considered the following search strategies: random-search, as well as Bayesian-based [43] and Hyperband-based [44] algorithms. In preliminary experiments, we tested all

Table 1. Optimized hyperparameters, and hyperparameter search spaces. The rightmost column shows optimal hyperparameters evaluated with Hyperband-based tuning.

Layer	Hyperparameter	Search space	Final
Convolutional layer 1	Kernel length	25,35,50,75,125,175	35
	Number of filters	8,16,32,64,128	8
Convolutional layer 2	Kernel length	25,35,50,75,125,175	175
	Number of filters	8,16,32,64,128	128
Convolutional layer 3	Kernel length	25,35,50,75,125,175	175
	Number of filters	8,16,32,64,128	16
MaxPooling layer	Window/stride size	3,5,7,9	7
Dense layer	Number of nodes	16,32,64,128,256	64
Convolutional layer dropout	Dropout rate	0 to 0.6	0.1
Dense layer dropout	Dropout rate	0 to 0.6	0.0
Optimizer	Learning rate	0.0001 to 0.1	0.00163

https://doi.org/10.1371/journal.pone.0272167.t001

approaches, finding that the Hyperband-based approach outperformed the others. The search space and the selected hyperparameters are listed in Table 1. We adjusted Hyperband such that the algorithm was repeated five times, with one-third of hyperparameter configurations kept during each successive halving sub-operation. To make hyperparameter optimization computationally tractable, some hyperparameters (e.g., number of convolutional layers and activation function) were selected based on previous work, specifically Schirrmeister and colleagues [27].

Evaluation process

For our network, we performed subjectwise 10-fold cross-validation to assess performance. For each of these 10 folds, subjects were allocated to either the training set (approx. 81%), the testing set (approx. 10%), or the validation set (approx. 9%). The training and validation sets were randomly undersampled based on the minority class for each subject's recording. Undersampling was not performed on the testing set (Table 2). We reasoned that training on highly unbalanced data is undesirable, as the classifier may develop bias towards the majority class (i.e., the non-apnea class). Network training was performed for 40 epochs (i.e., 40 passes of the training data through the network) with Python 3.7 and Tensorflow 2.2, on NeSI (New Zealand eScience Infrastructure), a high-performance computing platform which uses Tesla P100 GPU cards (NVIDIA, Santa Clara, California, United States). Computation took on average 112 microseconds per input sample for training and 40 microseconds per sample for testing. Furthermore, to evaluate our CNN's ability to generalize across datasets, we tested its performance (after training it using the SHHS dataset) on the UCD and MIT-BIH datasets described above.

To quantify performance of our trained CNN, we used accuracy and Matthews correlation coefficient (MCC). Accuracy is a common metric used to evaluate neural network

Table 2. The mean distribution of annotations within the training and testing sets. The validation set has the same proportions as the training set. Each EEG segment has an apnea annotation (i.e., "apnea" or "non-apnea") and a sleep-stage annotation (i.e., "wake", "REM" or "NREM"). Overall there was on average 1,144 segments per patient before undersampling and 378 segments per patient after resampling.

Number of annotations	Apnea	Non-apnea	Wake	REM	NREM
Training	407625.6	407625.6	202858.8	174796.6	437595.8
Testing	250424.2	50461.9	118989.6	35089.1	146807.4

https://doi.org/10.1371/journal.pone.0272167.t002 performance. However, accuracy may be susceptible to biases when data sets are unbalanced (i.e., data sets containing a preponderance of one or other class labels). Therefore, we also calculated MCC—a robust performance measure suitable for unbalanced testing data sets [45, 46]. Additionally, we developed a shuffle test to conservatively estimate performance baselines (referred to hereafter as the "conservative baseline"). To estimate baselines, we shuffled the class labels in the training set and performed training and testing using 10-fold subjectwise cross-validation; the measured performance provided our baselines. We reasoned that a CNN trained on these shuffled labels would be incapable of learning salient EEG features for SA detection, but could nonetheless learn the statistics of data set imbalance, and bias its behaviour accordingly. We used a Bayesian t-test, computing 95% highest density intervals (HDIs) [47], to compare our CNN's performance to baseline.

Critical-band masking

To explain the mechanisms of our trained networks, we used a critical-band masking (CBM) technique that was adapted from psychoacoustics [31] and visual psychophysics [32]. Here, we added bandlimited noise to test segments (but not training segments). We used a noise bandwidth of 1.5 Hz and parametrically varied the noise frequency centred from 1.5 Hz to 60 Hz in increments of 1.5 Hz. A finite-impulse-response (FIR) band-pass filter was applied to white noise to create this bandlimited noise (length = 825, transitional bandwidth = 0.5 Hz). At each center frequency, we quantified noise intensity by computing the log of the noise root-mean-square (RMS) value and signal RMS value to find the signal-to-noise ratio (SNR). For each band center frequency, we tested the trained network using these noisy test segments, quantifying the deleterious effect of noise by observing changes in MCC scores. This CBM process was performed for every fold of our subjectwise 10-fold cross-validation (see Evaluation Process).

Filter lesioning

We adapted a "lesioning" technique from Lawhern et al. [33] to determine the relative importance of first-layer convolutional kernels to the trained network. On each fold, after having trained the network, we zeroed all coefficients for one convolutional kernel, and then tested the network. We therefore quantified the deleterious effect that zeroing (i.e., lesioning) kernels had on test performance. We did this for each first-layer convolutional kernel on each fold of our subjectwise 10-fold cross-validation. Thus, on each fold, we were able to rank 1st-layer convolutional kernels by importance; e.g., the most important kernel, when lesioned, caused the greatest reduction in network performance. To verify that this lesioning technique was effective in identifying important convolutional kernels, we computed a correlation coefficient between all pairs of kernels within and between folds; specifically, we computed Pearson's correlation coefficient between kernels' Fourier amplitude spectra. The correlation coefficient was generally higher between convolutional kernels deemed to be important, both within and between folds. Finally, we computed the Fourier transform, and calculated the z-scores of the most important kernels (i.e., those determined important by Filter Lesioning). To do so, we formed a null distribution of kernel transforms using all trained kernels across all folds.

Results

Our network detected SA with an accuracy equal to 69.9% (the mean across folds of our subjectwise 10-fold cross-validation). The standard deviation (s.t.d.) of this accuracy was 3.0 percentage points across folds. Our network performed with an MCC score equal to 0.375 (s.t.d. = 0.017). The MCC performance metric was reliably above the conservative baseline (difference

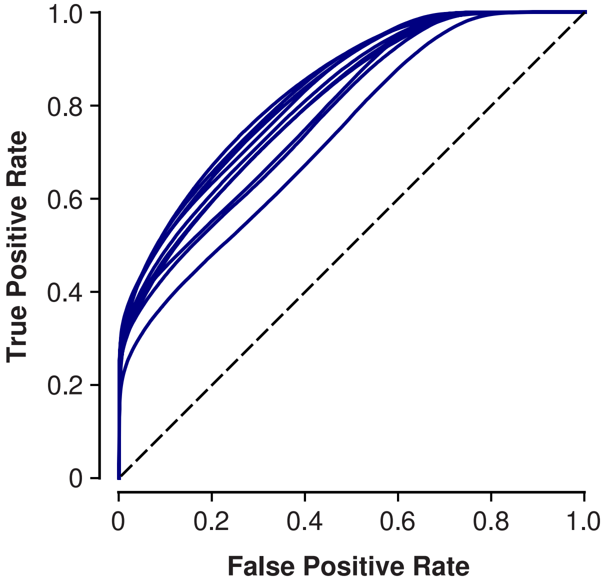

Fig 3. ROC curves summarizing the performance of our SA network. We represent each fold of our subjectwise 10-fold cross-validation with a separate curve. Across folds, the AUC averaged 0.804 (s.t.d. = 0.031).

https://doi.org/10.1371/journal.pone.0272167.g003

in MCC: Bayesian t-test, 0.374, 100% HDI; Methods); baseline MCC was 0.00 (s.t.d. = 0.04). Baseline accuracy was 49.9% (s.t.d. = 32.4 percentage points). After training with shuffled labels, the network tended to guess, from one fold to the next, either all apnea or all non-apnea. This resulted in a bimodal distribution for baseline accuracy, which we therefore did not use with a Bayesian t-test for statistical inference. In addition to measuring accuracy and MCC, we also undertook a receiver operating characteristic (ROC) analysis of our network's performance (Fig 3). Averaged across folds, the area under the ROC curve (AUC) was equal to 0.804 (s.t.d. = 0.031).

We then tested this trained network using data drawn from other datasets. Our network predicted the MIT dataset with a MCC equal to 0.216 (s.t.d. = 0.125). These performance metrics were reliably above our conservative baseline with a probability of 95.6% (difference in MCC: Bayesian t-test, 0.201); baseline MCC was -0.03 (s.t.d. = 0.18). Our network predicted the UCD dataset with a MCC equal to 0.169 (s.t.d. = 0.120). The performance metrics after testing on the UCD dataset were reliably above our conservative baseline (difference in MCC: Bayesian t-test, 0.157, 100% HDI; Methods); baseline MCC was 0.025 (s.t.d. = 0.09).

Several previous studies have found that SA is associated with sleep stage (a point we elaborate in Discussion). Therefore, we wondered if our trained SA network, to aid its classification, was representing sleep stage-associated features (i.e., covertly decoding sleep stage). To examine this idea, we decomposed the confusion matrix (Fig 4) into three submatrices, each corresponding to one of three sleep stages (Fig 5): wake, rapid-eye-movement (REM), and non-rapid-eye-movement (NREM). Upon inspection, parts of these submatrices appeared to indicate our SA network behaved in a biassed fashion. For instance, segments recorded during REM sleep predicted apnea on 93% of all trials. This potentially indicates that our trained SA network was representing (i.e., covertly decoding) sleep stage and, since SA is associated with sleep stage, used these representations to aid its performance in SA detection.

To test this hypothesis—that our CNN was representing sleep stage—we adapted our shuffle test (see Evaluation Procedure) to account for sleep stage; training data were split according to three sleep stages (wake, REM, NREM) and, for each stage, the labels "apnea" and "non-

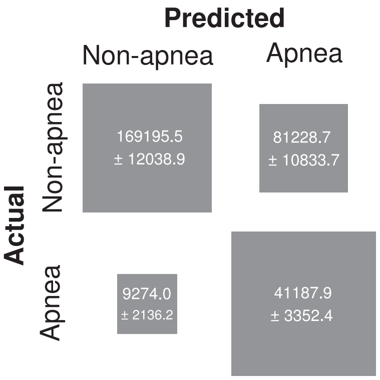

Fig 4. Confusion matrix, summarizing the performance of our SA network. The area of each square represents the value of each matrix entry. Values are counts averaged across our subjectwise 10-fold cross-validation. The intervals (±) associated with each value show s.t.d. across folds. Overall, the network performed with accuracy = 76.8%, as indicated by the mass along the matrix's main diagonal.

https://doi.org/10.1371/journal.pone.0272167.g004

apnea" were shuffled prior to re-training the network (below, we refer to this procedure as re-training the network after a "stage-wise shuffle" of the data). We reasoned that, if our SA network was, in fact, representing sleep stage to aid its detection of SA, this stage-wise shuffle would isolate the effect of this representation on SA detection performance. Specifically, if, after we re-trained the network using stage-wise shuffled data, network performance was unaltered, then that would indicate that the representation of sleep stage was wholly responsible for SA detection. On the other hand, if, after we re-trained the network using stage-wise shuffled data, network performance fell to baseline, then that would indicate that the representation of sleep stage was not being used to aid SA detection. After re-training the network on stage-wise shuffled data, the network performed with an accuracy equal to 56.7% (the mean across folds of our subjectwise 10-fold cross-validation; s.t.d. = 4.1 percentage points across folds), and a MCC score equal to 0.275 (s.t.d. = 0.019). The MCC was reliably below that of the original SA network (difference MCC: Bayesian t-test, -0.099, 100% HDI; Methods). Therefore, our SA network appeared to learn a representation of sleep stage to aid its detection of SA, but the network's learning of this representation only partly accounted for its ability to detect SA.

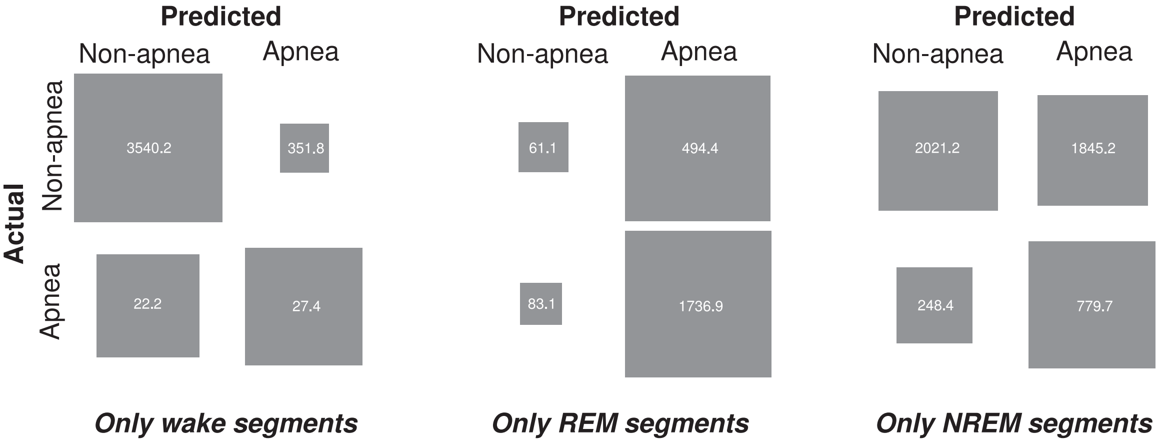

Fig 5. Confusion submatrices, each corresponding to one or other of three sleep stages: Wake (left), REM (middle), and NREM (right). For wake and REM, our SA network appeared to behave in a biased fashion. Graphical conventions are as in Fig 4.

https://doi.org/10.1371/journal.pone.0272167.g005

To explore further this idea—that our SA network was representing sleep stages—we developed a second CNN. This second CNN was nearly identical to our original SA network (the one exception being that the second CNN had five nodes in the final dense layer as opposed to two); we trained this network to decode sleep stages (wake, REM, and NREM). This sleep-stage network performed with an accuracy equal to 85.3% (mean across folds of our subject-wise 10-fold cross-validation; s.t.d. = 0.57 percentage points across folds), and a MCC score equal to 0.766 (s.t.d. = 0.009 across folds). The MCC was reliably below that of the original SA network (difference in MCC: Bayesian t-test, 0.766, 100% HDI; Methods); baseline accuracy for our sleep stage network was 52.0% (s.t.d. = 0.54 percentage points), and baseline MCC was 0.000 (s.t.d. = 0.001). Therefore, a network with architecture nearly identical to that of our SA network can be trained, explicitly, to decode sleep stage. This adds further support to our idea that our SA network learnt to represent (i.e., covertly decode) sleep stage, and it used this representation to aid in the detection of SA.

We wondered what EEG features were used by our SA network in performing SA detection. Therefore, we used critical-band masking (CBM), adding bandlimited noise to the signals used to test our SA network (Methods). The effects of CBM were graded; for the addition of low-intensity noise (SNR = 20; Methods), masking had little effect on network performance, regardless of the noise band's center frequency. When we increased the intensity of noise, network performance deteriorated (i.e., MCC decreased). Deterioration was pronounced for noise in some frequency bands but not others. Overall, the effects of CBM were primarily limited to three regions (Fig 6A): frequencies less than 4 Hz (the delta band); 30 to 45 Hz (the gamma band); and frequencies running from approximately 10 to 20 Hz, encompassing alpha (8 to 13 Hz), sleep spindles (11 to 16 Hz), and the lower beta band (14 to 30 Hz). Notably, when high-intensity (SNR = 0) noise was added to the delta band, MCC for SA detection was reduced from approximately 0.38 to 0.14. When high-intensity noise was added to the band associated with sleep spindles (11 to 16 Hz), MCC for SA detection was reduced from approximately 0.38 to 0.16. When high-intensity noise was added to the gamma band, MCC for SA detection was reduced from approximately 0.38 to a minimum of 0.22.

For comparison, we also applied CBM to the network that we re-trained using stage-wise shuffled data (see above). We reasoned that the different effects of CBM on those two networks (our original SA network, and the network re-trained on stage-wise shuffled data) would help determine which frequency bands were important to SA detection per se, which were important to sleep stage decoding (which appears to play a role in SA detection), and which were important to both. A key outcome of this experiment involved the delta band (Fig 6B); while noise in the delta band caused deterioration in the performance of our original SA network, it had relatively little effect on the network re-trained on stage-wise shuffled data. This difference indicated that, while other frequency bands may have contributed to the network's representation of sleep stage (e.g., sleep spindles, 11 to 16 Hz), delta band activity was specifically important to SA detection. This result is consistent with several studies that have observed an association between SA events during sleep and delta-band activity [27, 34] (a point elaborated in Discussion).

Critical band masking indicated that specific frequency bands were important for SA detection, especially the delta band. We therefore wondered whether filters in the first convolutional layer of our trained SA network responded selectively to EEG signals in these bands. To answer this question, we lesioned first-layer filters comprising the trained SA network (Methods). We found that, when lesioned, some filters substantially reduced the performance of our SA network (i.e., these filters appeared important to the network). The lesioning of other filters, by comparison, caused negligible deleterious effects (i.e., apparently unimportant filters). We sorted 1st-layer filters based on importance. These sorted filters were then used to generate

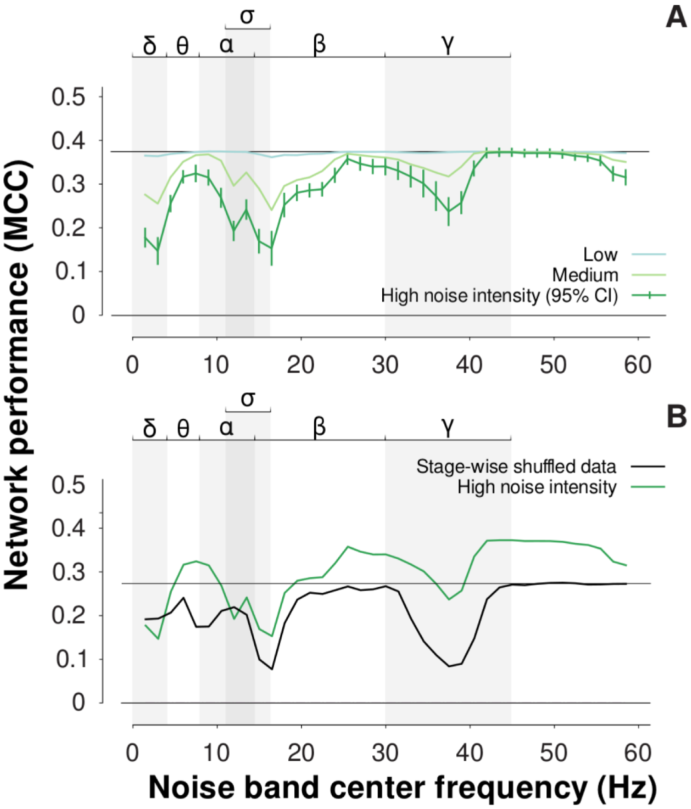

Fig 6. (A) Effect of critical-band masking on our SA network's performance. We used high-, medium-, and low-intensity noise: SNR = 5, 10, and 20, respectively (Methods). Overall, high-intensity noise decreased performance more than low-intensity noise. The deleterious effect of noise was pronounced in some frequency bands but not others. E.g., adding bandlimited noise to test signals in the delta band (< 4 Hz) caused MCC to decrease from 0.38 to 0.14. The lower, horizontal solid line (performance = 0.0 MCC) indicates the performance baseline (Methods), and the upper, horizontal solid line indicates network performance in the absence of noise (MCC = 0.38). The error bars (shown only for high-intensity noise) are 95% confidence intervals computed across folds of our subjectwise 10-fold cross-validation. The Greek letters (top) mark traditional frequency bands; sigma marks the band associated with sleep spindles. (B) Effect of critical-band masking on our re-trained SA network; we re-trained the network after data were stage-wise shuffled. Adding bandlimited noise to test signals in the delta and alpha (8 to 13 Hz) bands, here, had little effect on network performance. By contrast, noise in the lower beta band and gamma band heavily reduced performance. Other graphical conventions are as in (A). Upper horizontal line marks the stage-wise shuffled no-noise response (MCC = 0.275).

https://doi.org/10.1371/journal.pone.0272167.g006

an "importance matrix" (Fig 7). In light of that importance matrix we asked, "Do important filters have similar characteristics?" To answer this question, we computed each 1st-layer filter's amplitude spectrum using the Fourier transform, and we averaged spectra within importance and across folds (i.e., after Fourier transformation, we averaged amplitude spectra within rows, across columns, of Fig 7). When we compared the amplitude spectra of relatively important filters to the ensemble (i.e., all filters comprising Fig 7), we found the following (Fig 8): important 1st-layer filters tended to attenuate the delta band, and amplify the beta and gamma bands. This pattern is consistent with the result of our critical-band masking.

Discussion

We have developed a CNN that detects SA from single-channel EEG (accuracy = 69.9%; baseline = 49.9%). Our CNN's performance, when measured using MCC (MCC = 0.38; baseline = 0.03), showed a "weak-moderate positive correlation" [48] (conventionally, MCC is

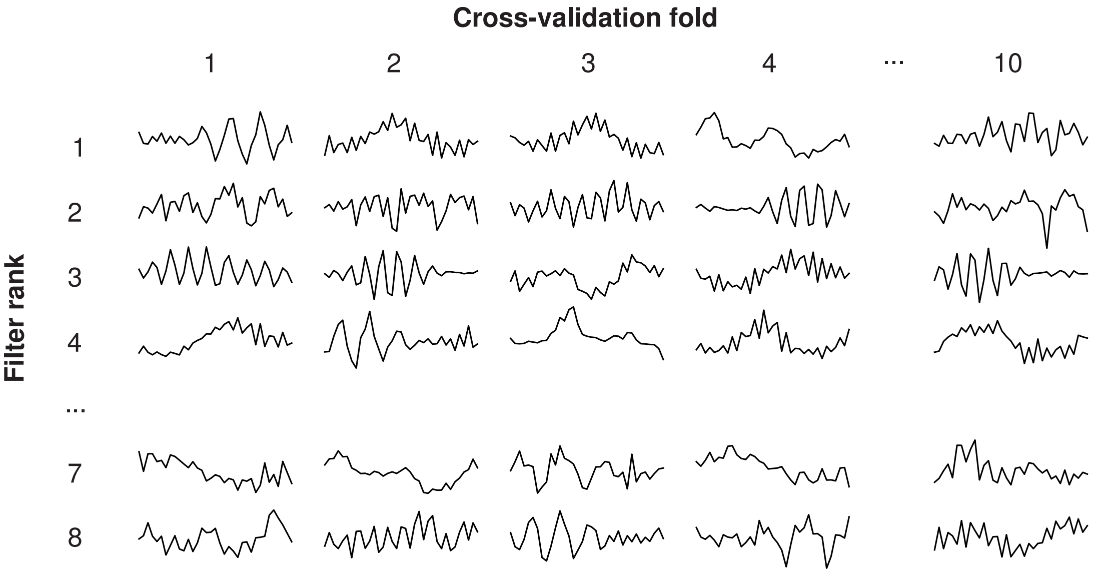

Fig 7. Importance matrix, showing 1st-layer filters comprising our trained SA network. Matrix columns correspond to folds from our subjectwise 10-fold cross-validation; rows correspond to importance (i.e., the most important filter on each fold is shown in row 1). To illustrate by example, on the first fold of cross-validation, the filter kernel illustrated at column 1, row 1 (top-left), was determined to be the most important; lesioning this filter reduced the trained SA network's performance more significantly than any other filter on this fold.

https://doi.org/10.1371/journal.pone.0272167.g007

interpreted in a similar fashion to Pearson's correlation coefficient [48]). Furthermore, our network's performance was robust when measured using ROC analysis (across folds, average AUC = 0.804). To explain the mechanisms of our CNN, we used two techniques—critical-band masking (wherein band-limited noise was added to signals used to test the trained network) and filter lesioning (wherein 1st-layer filter kernels were zeroed in turn). To our

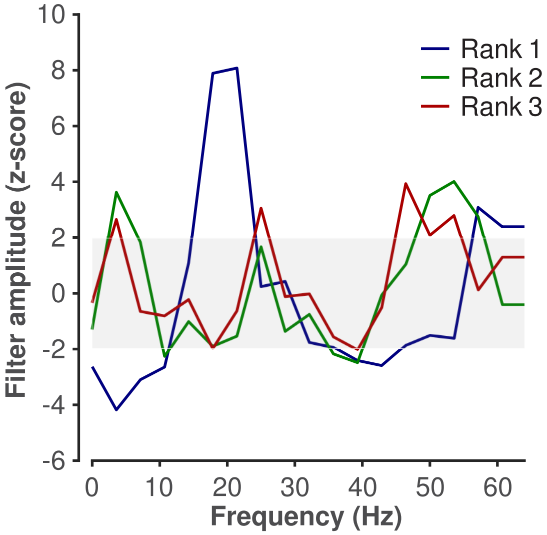

Fig 8. Amplitude spectra of 1st-layer filters important to the SA network's performance. We show spectra for the 1st-, 2nd-, and 3rd-most important filters ("Rank 1, 2, and 3", respectively). Important filters appeared to attenuate the delta band, and amplify the beta and gamma bands. The shaded rectangle marks a 95%-confidence interval (i.e., -1.96 < z-score < 1.96), wherein the spectral amplitude of rank 1, 2, and 3 filters is not appreciably different from that of all other 1st-layer filters comprising the ensemble.

https://doi.org/10.1371/journal.pone.0272167.g008

knowledge, the use of critical band masking is a novel approach to analysing a trained CNN. Our results indicate all three of our hypotheses, which we outlined in the Introduction, were confirmed: there is information in single-channel EEG enabling reliable subjectwise detection of SA by an explainable CNN; knowledge of sleep stage appeared to improve SA detection; and, our CNN used information contained in single-channel EEG that is consistent with known SA biomarkers, specifically the delta and beta bands.

Previous work has demonstrated that SA is accompanied by characteristic alterations of EEG [49, 50]. Likewise, the sleep stages—wake, REM, and the three NREM stages (N1, N2, N3)—all also are accompanied by systematic changes in EEG [51, 52]. Azim and colleagues [50] studied the normalized Welch power spectral density of electrode C4-A1 from the UCD database [36]. They found that during SA events, power in the beta band (i.e., frequencies between 14 and 30 Hz) decreased (compared to pre-apnea events), before rising again after SA event termination. During different stages of sleep, the EEG features associated with SA can change. For example, apneas occurring during NREM sleep are associated with a gradual increase in delta-band activity (i.e., at frequencies $<$ 4 Hz), followed by a decrease within that band concomitant with patient arousal and/or wakefulness [49, 50]. In contrast, apneas occurring during REM sleep are associated with transient increases in delta-band activity [49, 50]. REM SA events are also generally associated with small increases to beta band activities. Physiological differences between sleep stages can also impact SA. REM sleep causes the relaxation of muscle tone which is conducive to SA events [53]. Therefore, EEG—a signal which contains characteristics for both SA and sleep stage—should in theory be useful for detecting SA.

Our use of critical-band masking and filter lesioning, taken together, indicated that our SA network learned to rely on known SA as well as sleep stage biomarkers. Critical-band masking indicated that delta-band activity was important to the detection of SA; that beta- and gamma-band activity was important to the decoding of sleep stage; and that alpha-band activity may have played a role in both SA detection and sleep stage decoding. We made these inferences by adding noise, first, to signals used to test our trained SA network, and, second, to signals used to test the network after re-training with stage-wise shuffled data. Our masking results were in broad agreement with our lesioning results. The trained SA network's most important 1st-layer filters selectively attenuated delta-band activity, and selectively amplified activity between 14 and 30 Hz (i.e., the beta band). Taken together, our findings are broadly consistent with existing work; it has been previously shown using spectral analysis of multi-channel EEG recorded from SA sufferers that delta-band activity is associated with SA events [49, 50]. Furthermore, REM and NREM sleep are associated with the reduction of power in beta- and gamma bands [54].

The present study is subject to two main limitations, the first of which concerns the generalisation of our CNN across datasets. We trained our CNN using the SHHS dataset; subjectwise 10-fold cross-validation using test data from SHHS (ie., within-dataset testing) showed good performance (see Results). However, when we tested our SHHS-trained CNN using data from the UCD or MIT-BIH datasets (ie., across-dataset testing), its performance was somewhat sobering (see Results). This performance reduction is possibly attributable to low inter-rater reliability (IRR). For the SHHS dataset, the three main SA scorers were in only moderate agreement (scorers 914 and 915, Cohen's kappa = 0.7; scorers 914 and 916, 0.73; scorers 915 and 916, 0.76) [55]. Estimates of IRR between datasets (e.g., between SHHS and UCD) is presently unavailable, however it stands to reason that it is significantly lower than 0.7 because SA annotation procedures differed considerably between datasets (see Methods). The precise relationship between IRR and the theoretical maximum performance of our CNN, and the way in which changes to IRR in turn affect CNN performance, is a subject of our ongoing research. The second limitation of our study concerns labelling; we labelled segments used in the training, testing, and validation of our CNN as either "apnea" or "non-apnea" despite the fact that there are several clinical SA subtypes, including obstructive, central, and mixed SA, and hypopnea [56]. We grouped these subtypes for two reasons. First, there were relatively few instances of central SA in these datasets; had we attempted to balance our training set, for each subject, with regard to subtype, there would have been too few data to train a reliable CNN. Second, IRR is low as regards discriminating obstructive from mixed SA; in the SHHS dataset, these subtypes are grouped under the label "obstructive SA". The development of an architecture capable of discriminating between SA subtypes is a subject of our ongoing research.

To our knowledge, Jiang and colleagues [57] is the only other group to develop a CNN for SA classification using single-channel EEG. We have extended their work in several ways, albeit comparison between their results and ours, for reasons outlined below, is not straightforward. First, Jiang et al. used a small dataset (the MIT-BIH Polysomnographic Database [37]), comprising recordings from only 16 participants. By contrast, our study integrated recordings from three databases, comprising over 2600 participants. Second, Jiang et al. appear to have used an unbalanced data set to train and test their network (specifically, their dataset appears to have overrepresented apnea, as opposed to non-apnea, EEG recordings); in general, the use of unbalanced data sets may bias estimates of a classifier's performance. By contrast, we were careful to balance data before training our CNN, and we developed a shuffle test to ensure that our estimates of baseline performance were conservative. Third, Jiang et al. performed pooled (not subjectwise) cross-validation. By "pooled", we mean that on each fold of their cross-validation procedure, recordings from all subjects were contained in both training and testing sets. Because they pooled data, it is unclear whether their results can generalize in a clinical setting (i.e., it would be clinically useful if a CNN, trained using data recorded from a normative cohort of subjects 1 through N, could be used to detect SA in recordings from a previously unseen subject N+1). By contrast, we used subjectwise cross-validation, which indicates that our results will generalize to unseen participants, and therefore may have clinical potential for diagnosis. Fourth, the network of Jiang et al. was, by contrast to ours, rather complex, and therefore ill-suited for explanation; the network employed a hybrid architecture with 4 parallel, computational branches: two shallow branches (each of which comprised three convolutional layers), and two deep branches (comprising 6 and 9 convolutional layers, respectively). By contrast, our network was simple, and this simplicity enabled an explanation of its function. It is therefore our expectation that our explainable CNN has clinical potential for the improved detection, diagnosis, and treatment of SA.

Conclusion

EEG offers an alternative way to detect SA, and detection appears to benefit from sleep stage information contained in EEG. Single-channel EEG is low-instrumentation, making it potentially suitable for in-home sleep and SA monitoring. Here, we have developed a CNN that reliably detects SA from single-channel EEG. Our visualization techniques indicated that the classifier learned frequency-band information consistent with known SA biomarkers. It is a priority for future work to quantify our algorithm on consumer-grade EEG equipment.

Acknowledgments

We thank Hazel Glasgow and Lucia Quirke for their valuable contributions at an early stage of this research. We used high-performance computing facilities provided by New Zealand eScience Infrastructure (https://www.nesi.org.nz), which is jointly funded by collaborator institutions and the New Zealand Ministry of Business, Innovation and Employment.

Author Contributions

Conceptualization: Lachlan D. Barnes, Kevin Lee, Andreas W. Kempa-Liehr, Luke E. Hallum.

Data curation: Lachlan D. Barnes, Kevin Lee, Andreas W. Kempa-Liehr, Luke E. Hallum.

Formal analysis: Lachlan D. Barnes, Kevin Lee, Andreas W. Kempa-Liehr, Luke E. Hallum.

Funding acquisition: Lachlan D. Barnes, Kevin Lee, Andreas W. Kempa-Liehr, Luke E. Hallum.

Investigation: Lachlan D. Barnes, Kevin Lee, Andreas W. Kempa-Liehr, Luke E. Hallum.

Methodology: Lachlan D. Barnes, Kevin Lee, Andreas W. Kempa-Liehr, Luke E. Hallum.

Project administration: Lachlan D. Barnes, Kevin Lee, Andreas W. Kempa-Liehr, Luke E. Hallum.

Resources: Lachlan D. Barnes, Kevin Lee, Andreas W. Kempa-Liehr, Luke E. Hallum.

Software: Lachlan D. Barnes, Kevin Lee, Andreas W. Kempa-Liehr, Luke E. Hallum.

Supervision: Lachlan D. Barnes, Kevin Lee, Andreas W. Kempa-Liehr, Luke E. Hallum.

Validation: Lachlan D. Barnes, Kevin Lee, Andreas W. Kempa-Liehr, Luke E. Hallum.

Visualization: Lachlan D. Barnes, Kevin Lee, Andreas W. Kempa-Liehr, Luke E. Hallum.

Writing – original draft: Lachlan D. Barnes, Kevin Lee, Andreas W. Kempa-Liehr, Luke E. Hallum.

Writing – review & editing: Lachlan D. Barnes, Kevin Lee, Andreas W. Kempa-Liehr, Luke E. Hallum.

References

1. Jennum P, Riha RL. Epidemiology of sleep apnoea/hypopnoea syndrome and sleep-disordered breathing. Eur Resp J. 2009 Apr; 33(4):907–14. https://doi.org/10.1183/09031936.00180108 PMID: 19336593
2. Punjabi NM. The epidemiology of adult obstructive sleep apnea. Proc Am Thorac Soc. 2008 Feb 15; 5 (2):136–43. https://doi.org/10.1513/pats.200709-155MG PMID: 18250205
3. Ulfberg J, Carter N, Edling C. Sleep-disordered breathing and occupational accidents. Scand J Work Environ Health. 2000 Jun; 26(3):237–42. https://doi.org/10.5271/sjweh.537 PMID: 10901116
4. Sullivan CE, Issa FG, Berthon-Jones M, Eves L. Reversal of obstructive sleep apnoea by continuous positive airway pressure applied through the nares. Lancet. 1981 Apr 18; 1(8225):862–5. https://doi.org/10.1016/S0140-6736(81)92140-1 PMID: 6112294
5. Marin JM, Carrizo SJ, Vicente E, Agusti AG. Long-term cardiovascular outcomes in men with obstructive sleep apnoea-hypopnoea with or without treatment with continuous positive airway pressure: an observational study. Lancet. 2005 Mar 19–25; 365(9464):1046–53. https://doi.org/10.1016/S0140-6736(05)71141-7 PMID: 15781100
6. Rundo JV, Downey R III. Polysomnography In: Levin KH, Chauvel P (eds) Handbook of Clinical Neurology, vol. 160, pp. 381–392. New York: Elsevier; 2019.
7. Park KS, Choi SH. Smart technologies toward sleep monitoring at home. Biomed Eng Lett. 2019 Jan 11; 9(1):73–85. https://doi.org/10.1007/s13534-018-0091-2 PMID: 30956881
8. Flemons WW, Douglas NJ, Kuna ST, Rodenstein DO, Wheatley J. Access to diagnosis and treatment of patients with suspected sleep apnea. Am J Respir Crit Care Med. 2004 Mar 15; 169(6):668–72. https://doi.org/10.1164/rccm.200308-1124PP PMID: 15003950
9. Young T, Evans L, Finn L, Palta M. Estimation of the clinically diagnosed proportion of sleep apnea syndrome in middle-aged men and women. Sleep. 1997 Sep; 20(9):705–6. https://doi.org/10.1093/sleep/20.9.705 PMID: 9406321
10. Mykytyn IJ, Sajkov D, Neill AM, McEvoy RD. Portable computerized polysomnography in attended and unattended settings. Chest. 1999 Jan; 115(1):114–22. https://doi.org/10.1378/chest.115.1.114 PMID: 9925071

11. Bruyneel M, Sanida C, Art G, Libert W, Cuvelier L, Paesmans M, et al. Sleep efficiency during sleep studies: results of a prospective study comparing home-based and in-hospital polysomnography. J Sleep Res. 2011 Mar; 20(1 Pt 2):201–6. https://doi.org/10.1111/j.1365-2869.2010.00859.x PMID: 20561176

12. Mikkelsen KB, Villadsen DB, Otto M, Kidmose P. Automatic sleep staging using ear-EEG. Biomed Eng Online. 2017 Sep 19; 16(1):111. https://doi.org/10.1186/s12938-017-0400-5 PMID: 28927417

13. Holzinger A. Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inform. 2016 Jun; 3(2):119–131. https://doi.org/10.1007/s40708-016-0042-6 PMID: 27747607

14. Mostafa SS, Mendonça F, Ravelo-García AG, Morgado-Dias F. A systematic review of detecting sleep apnea using deep learning. Sensors. 2019 Nov 12; 19(22):4934. https://doi.org/10.3390/s19224934 PMID: 31726771

15. Varon C, Caicedo A, Testelmans D, Buyse B, Van Huffel S. A novel algorithm for the automatic detection of sleep apnea from single-lead ECG. IEEE Trans Biomed Eng. 2015 Sep; 62(9):2269–2278. https://doi.org/10.1109/TBME.2015.2422378 PMID: 25879836

16. Wang T, Lu C, Shen G, Hong F. Sleep apnea detection from a single-lead ECG signal with automatic feature-extraction through a modified LeNet-5 convolutional neural network. PeerJ. 2019 Sep 20; 7: e7731. https://doi.org/10.7717/peerj.7731 PMID: 31579607

17. Mostafa SS, Mendonca F, Morgado-Dias F, Ravelo-Garcia A. SpO2 based sleep apnea detection using deep learning. IEEE 21st International Conference on Intelligent Engineering Systems (INES), pp. 91–96, 2017.

18. Kim J, ElMoaqet H, Tilbury DM, Ramachandran SK, Penzel T. Time domain characterization for sleep apnea in oronasal airflow signal: a dynamic threshold classification approach. Physiol Meas. 2019 Jun 4; 40(5):054007. https://doi.org/10.1088/1361-6579/aaf4a9 PMID: 30524019

19. ElMoaqet H, Kim J, Tilbury D, Ramachandran SK, Ryalat M, Chu CH. Gaussian mixture models for detecting sleep apnea events using single oronasal airflow record. Appl Sci-Basel. 2020; 10(21):7889. https://doi.org/10.3390/app10217889

20. ElMoaqet H, Eid M, Glos M, Ryalat M, Penzel T. Deep recurrent neural networks for automatic detection of sleep apnea from single channel respiration signals. Sensors. 2020 Sep 4; 20(18):5037. https://doi.org/10.3390/s20185037 PMID: 32899819

21. Liu D, Pang Z, Lloyd SR. A neural network method for detection of obstructive sleep apnea and narcolepsy based on pupil size and EEG. IEEE Trans Neural Netw. 2008 Feb; 19(2):308–18. https://doi.org/10.1109/TNN.2007.908634 PMID: 18269961

22. Casson A, Yates D, Smith S, Duncan J, Rodriguez-Villegas E. Wearable electroencephalography. IEEE Eng Med Biol Mag. 2010; 29(3): 44–56. https://doi.org/10.1109/MEMB.2010.936545 PMID: 20659857

23. Almuhammadi WS, Aboalayon KAI, Faezipour M. Efficient obstructive sleep apnea classification based on EEG signals. IEEE Conference on Long Island Systems, Applications and Technology (LISAT), pp. 1–6, 2015.

24. Saha S, Bhattacharjee A, Fattah SA. Automatic detection of sleep apnea events based on inter-band energy ratio obtained from multi-band EEG signal. Healthc Technol Lett. 2019 Jun 3; 6(3):82–86. https://doi.org/10.1049/htl.2018.5101 PMID: 31341633

25. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May 28; 521(7553):436–44. https://doi.org/10.1038/nature14539 PMID: 26017442

26. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017 May 60:84–90. https://doi.org/10.1145/3065386

27. Schirrmeister RT, Springenberg JT, Fiederer LDJ, Glasstetter M, Eggensperger K, Tangermann M, et al. Deep learning with convolutional neural networks for EEG decoding and visualization. Hum Brain Mapp. 2017 Nov; 38(11):5391–5420. https://doi.org/10.1002/hbm.23730 PMID: 28782865

28. Casal R, Di Persia LE, Schlotthauer G. Temporal convolutional networks and transformers for classifying the sleep stage in awake or asleep using pulse oximetry signals. J Comput Sci. 2022; 59: 101544. https://doi.org/10.1016/j.jocs.2021.101544

29. Kundu S. AI in medicine must be explainable. Nat Med. 2021 Aug; 27(8):1328. https://doi.org/10.1038/s41591-021-01461-z PMID: 34326551

30. Barrett PM, Komatireddy R, Haaser S, Topol S, Sheard J, Encinas J, et al. Comparison of 24-hour Holter monitoring with 14-day novel adhesive patch electrocardiographic monitoring. Am J Med. 2014 Jan; 127(1):95.e11–7. https://doi.org/10.1016/j.amjmed.2013.10.003 PMID: 24384108

31. Greenwood DD. Auditory Masking and the Critical Band. J Acoust Soc Am. 1961 Apr; 33:484–502. https://doi.org/10.1121/1.1908699

32. Solomon JA, Pelli DG. The visual filter mediating letter identification. Nature. 1994 Jun 2; 369 (6479):395–7. https://doi.org/10.1038/369395a0 PMID: 8196766

33. Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ. EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces. J Neural Eng. 2018 Oct; 15(5):056013. https://doi.org/10.1088/1741-2552/aace8c PMID: 29932424

34. Zhang GQ, Cui L, Mueller R, Tao S, Kim M, Rueschman M, et al. The National Sleep Research Resource: towards a sleep data commons. J Am Med Inf Assoc. 2018 Oct 1; 25(10):1351–1358. https://doi.org/10.1093/jamia/ocy064 PMID: 29860441

35. Quan SF, Howard BV, Iber C, Kiley JP, Nieto FJ, O'Connor GT, et al. The Sleep Heart Health Study: design, rationale, and methods. Sleep. 1997 Dec; 20(12):1077–85. PMID: 9493915

36. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 2000 Jun 13; 101(23):E215–20. https://doi.org/10.1161/01.CIR.101.23.e215 PMID: 10851218

37. Ichimaru Y, Moody GB. Development of the polysomnographic database on CD-ROM. Psychiatry Clin Neurosci. 1999 Apr; 53(2):175–7. https://doi.org/10.1046/j.1440-1819.1999.00527.x PMID: 10459681

38. Hobson JA. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. Electroencephalogr Clin Neurophysiol. 1969; 26(6): 644. https://doi.org/10.1016/0013-4694(69)90021-2

39. The Ten Twenty Electrode System: International Federation of Societies for Electroencephalography and Clinical Neurophysiology. American Journal of EEG Technology. 1961 Mar; 1(1):13–19. https://doi.org/10.1080/00029238.1961.11080571

40. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014; 15:1929–1958.

41. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

42. Bridle JS. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In: Soulié FF, Hérault J (eds) Neurocomputing. Berlin, Heidelberg: Springer; 1990.

43. Mockus J. Bayesian approach to global optimization: Theory and applications. Berlin, Heidelberg: Springer; 1989.

44. Li L, Jamieson K, DeSalvo G, Rostamizadeh A, Talwalkar A. Hyperband: A novel bandit-based approach to hyperparameter optimization. arXiv:1603. 06560 [cs, stat], Jun. 2018.

45. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020 Jan 2; 21(1):6. https://doi.org/10.1186/s12864-019-6413-7 PMID: 31898477

46. Delgado R, Tibau XA. Why Cohen's Kappa should be avoided as performance measure in classification. PLoS One. 2019 Sep 26; 14(9):e0222916. https://doi.org/10.1371/journal.pone.0222916 PMID: 31557204

47. Kruschke JK. Bayesian estimation supersedes the t test. J Exp Psychol-Gen. 2013 May; 142(2):573–603. https://doi.org/10.1037/a0029146 PMID: 22774788

48. Schober P, Boer C, Schwarte LA. Correlation coefficients: appropriate use and interpretation. Anesth Analg. 2018 May; 126(5):1763–1768. https://doi.org/10.1213/ANE.0000000000002864 PMID: 29481436

49. Svanborg E, Guilleminault C. EEG frequency changes during sleep apneas. Sleep. 1996 Apr; 19 (3):248–54. PMID: 8723384

50. Azim MR, Haque SA, Amin MS, Latif T. Analysis of EEG and EMG signals for detection of sleep disordered breathing events. IEEE International Conference on Electrical and Computer Engineering (ICECE), pp. 646–649, 2010.

51. Loomis AL, Harvey EN, Hobart GA. Cerebral states during sleep, as studied by human brain potentials. J Exp Psychol. 1937; 21(2):27–144. https://doi.org/10.1037/h0057431

52. Malhotra RK, Avidan AY. Sleep stages and scoring technique. In: Chokroverty S, Thomas RJ (eds) Atlas of sleep medicine (2nd edition). Elsevier; 2014.

53. Kandel ER, Schwartz JH, Jessell TM. Principles of neural science (4th edition). Stamford, CT: Appleton and Lange; 2000.

54. Gross DW, Gotman J. Correlation of high-frequency oscillations with the sleep-wake cycle and cognitive activity in humans. Neuroscience. 1999; 94(4):1005–18. https://doi.org/10.1016/S0306-4522(99)00343-7 PMID: 10625043

55. McHugh ML. Interrater reliability: the kappa statistic. Biochem Medica. 2012; 22(3):276–82. https://doi.org/10.11613/BM.2012.031 PMID: 23092060

56. Javaheri S, Barbe F, Campos-Rodriguez F, Dempsey JA, Khayat R, Javaheri S, et al. Sleep Apnea: Types, Mechanisms, and Clinical Cardiovascular Consequences. J Am Coll Cardiol. 2017 Feb 21; 69 (7):841–858. https://doi.org/10.1016/j.jacc.2016.11.069 PMID: 28209226

57. Jiang D, Ma Y, Wang Y. A multi-scale parallel convolutional neural network for automatic sleep apnea detection using single-channel EEG signals. IEEE 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pp. 1–5, 2018.