Behaviormetrika (2024) 51:341–358
https://doi.org/10.1007/s41237-023-00199-x

ORIGINAL PAPER

Generalisability of sleep stage classification based on interbeat intervals: validating three machine learning approaches on self-recorded test data

Stefan Kranzinger[1] · Sebastian Baron[2] · Christina Kranzinger[1] · Dominik Heib[2] · Christian Borgelt[2]

Received: 2 November 2022 / Accepted: 4 May 2023 / Published online: 18 May 2023

Abstract

Classifying sleep stages is an important basis for neuroscience, health sciences, psychology and many other fields. However, the manual determination of sleep stages is tedious and time consuming. Therefore, the development of automatic sleep stage classifiers based on data collected with low-cost sensor systems is an important research area. This study aims to analyse the generalisability of different machine learning approaches for sleep stage classification. We train three different models (random forest, CNN-LSTM and seq2seq) for classifying three as well as four sleep stages, with the MESA data set. For validation, we use a fivefold cross-validation and further validate the models with one new self-recorded test data set to analyse the models' generalisability to a completely new cohort with different characteristics with regard to age and health status. Our results show that the two deep learning approaches performed better than the random forest. Moreover, all models are generalisable and therefore suitable for sleep stage classification on a new three-stage classification data set. However, generalisability for the four-stage classification task shows poorer performance, and therefore requires new approaches such as transfer learning or a larger data set to train the models.

Keywords Deep learning · Random forest · Sleep stage classification · Classification

Communicated by Alfonso Iodice D'Enza.

Stefan Kranzinger
stefan.kranzinger@salzburgresearch.at

Christina Kranzinger
christina.kranzinger@salzburgresearch.at

[1] Salzburg Research Forschungsgesellschaft mbH, Jakob-Haringer-Straße 5/3, 5020 Salzburg, Austria

[2] Paris-Lodron-University Salzburg, Hellbrunner Straße 34, 5020 Salzburg, Austria

1 Introduction

Sleep plays a crucial role for our organism and is indispensable for our physical and emotional well-being. Consequently, nonrestorative sleep has been repeatedly linked to significant impairments in social, occupational or other areas of functioning causing massive socio-economic burdens. Alarmingly, general sleep disturbances are common, affecting about 1/3 of the adult general population (Kerkhof 2017; Chattu et al. 2019). In contrast, there is a considerable shortage of somnologists and qualified sleep laboratories causing unnecessary diagnosis delays. In addition, until today, sleep is measured on the basis of polysomnography (PSG) and classified by human experts into five different sleep stages according to the suggestion of the American Association for Sleep Medicine (AASM) Manual (Iber et al. 2007). PSG and manual sleep scoring is personnel intensive, time consuming and expensive. Therefore, new low-cost measurement technologies and automatised sleep scoring routines that allow sleep measurements in extensive field studies are needed to encounter the high prevalence of sleep problems in modern societies.

Interestingly recent studies (Radha et al. 2019; Sridhar et al. 2020; Sun et al. 2020) have shown that basic physiological signals, such as heart rate variability (HRV) and respiration frequency, substantially change over sleep allowing reliably classifying sleep solely on these signals into three to four sleep stages. So far, a wide range of classifiers have been used including deep learning models, support vector machines, random forests, bootstrap aggregation with a decision tree as the base learner, hidden Markov models or *k*-means clustering (Faust et al. 2019). For instance, Radha et al. (2019) applied long short-term memory (LSTM) neural networks on 132 HRV features. In addition, Sridhar et al. (2020) used a fully convolutional neural network (CNN) for sleep stage classification. In Zhai et al. (2020) the Multi-Ethnic Study of Atherosclerosis (MESA) database is used to compare random forests with CNN and LSTM models, where neural networks achieved higher classification accuracies than traditional machine learning models.

In addition to the development of various machine learning algorithms, it is of great interest that these models are able to classify completely new data sets to provide sleep stage classification without the collection of large amount of data and extensive model training. Loh et al. (2020) showed, in their systematic review, that in sleep stage classification, the majority of studies used data from only one database for model training and testing. Consequently, Loh et al. (2020) state that it is important to evaluate models across different databases as this could decrease the bias in the methods and identify the best performing approach. Sridhar et al. (2020) trained their algorithm with data from the Sleep Heart Health Study (SHHS) (Quan et al. 1997) and MESA and further validated it with the Physionet CinC (Ghassemi et al. 2018) data as an independent data set. Olesen et al. (2020) analysed a novel deep neural network with regard to generalisability and used five different data sources, containing subjects with diverse disease phenotypes. The authors find that automatic sleep stage classification should consider as much data from different sources as possible and recommend for future studies to test developed models on completely new cohorts. Jiang et al. (2019) used multimodal decomposition and hidden Markov model-based refinement based on a single-channel electroencephalography (EEG) of the Sleep-EDF (European Data Format) and Sleep-EDF expanded data sets. The authors further validated the generalisation ability of their model on the Montreal Archive of Sleep Studies database. Moreover, Guillot and Thorey (2021) used eight heterogeneous sleep staging data sets to train and validate a model based on leave-one-data-set-out. The three data sets MESA, MrOS and SHHS showed better generalisation than the other four smaller data sets which were used: Dreem Open Dataset - Obstructive, Dreem Open Dataset - Healthy, MASS, and Sleep-EDF. To address the need of low-cost measurements and to provide a low-threshold offer for sleep analysis, we have used low-cost measurement technologies to collect sleep data for which we aim to identify sleep stages using classification algorithms trained on freely available data sets. We compare three different algorithms (one classic machine learning approach and two deep learning models) and analyse the generalisability of machine learning approaches for sleep stage classification based on interbeat intervals (IBIs).

Therefore, we aim to answer the following two research questions:

1. How well does it work to classify sleep stages from self-recorded PSG data when using a model trained on a freely available database?
2. Which method generalises best to our data for three, as well as four sleep stage classification tasks?

To answer these research questions, we analyse whether our sleep stage classification models trained with the MESA data can be generalised to a new database with different characteristics, such as age distribution. In this context, we use three different classification approaches that have all already been used in the recent literature to classify sleep stages.[1] First, a random forest (Breiman 2001), as a tree-based and interpretable machine learning technique that has lower computational cost. Second, a CNN-LSTM, as a deep learning approach (Goodfellow et al. 2015; Hochreiter and Schmidhuber 1997). Third, as a second deep learning approach, we adopt a seq2seq model motivated by the architecture of Sridhar et al. (2020) that receives, in contrast to the random forest and the CNN-LSTM that are based on single time windows, the whole night for classification. We would like to emphasise that the main objective of this work is not to maximise the classification performance of the model, but to show which model generalises best on a completely new data set that consists of a different sample in terms of age and health status compared to the training data. Consequently, the results of this work show whether or which of these three approaches generalises better to our data, and hence promises better generalisability. We apply the classification task for two different scenarios. In the first, we classify

[1] For random forest and CNN-LSTM see, among others, Zhai et al. (2020) and for sequence to sequence (seq2seq) see, among others, Sridhar et al. (2020).

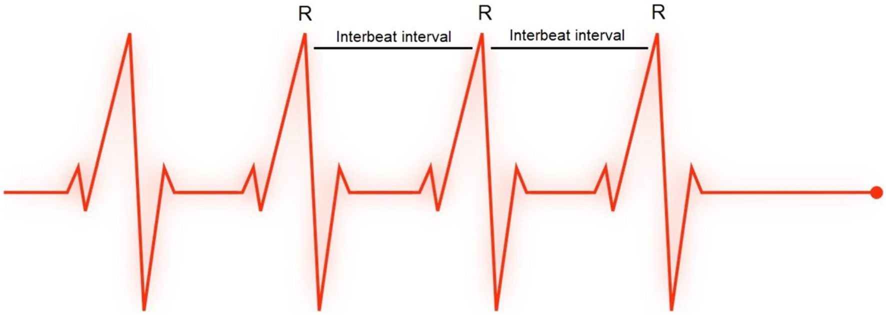

Fig. 1 Heart rate with R-peaks and IBIs

three sleep stages including *wake, NREM sleep* and *REM sleep*. In the second, we classify four sleep stages by dividing NREM into light and deep sleep and thus classify *wake, light sleep, deep sleep and REM sleep*.

2 Methods

In the following, we review the data used for this study and explain the feature engineering process in detail. We also present the algorithms used and their specification, as well as the evaluation method used to answer the two research questions.

2.1 Data

To classify the sleep stages, we use heart rate data that can be collected with low-cost sensor systems and that are based on the IBIs, the distance between two consecutive R-peaks of the heart rate. Figure 1 visualises a heart rate with the R-peaks and the IBIs.

For the training and validation of the sleep stage classifiers, different data sets were used: (i) MESA (Zhang et al. 2018; Chen et al. 2015) and (ii) self-recorded PSG data (Virtual Sleep Lab (VSL) data).

MESA: MESA is a collaborative longitudinal investigation which has been described previously by Bild (2002). 2237 participants of the overall MESA study sample were enrolled in a sleep exam that included a sleep questionnaire, a 7-day wrist actigraphy measuring activity intensity throughout the day and an in-home PSG. To make a final selection of the nights used for the analysis, we defined our own quality standard consisting of three sub-metrics. First, we use a combination of five quality metrics provided by MESA (wakslepr5, stg1stg2pr5, stg2stg3pr5, remnrempr5, arunrel5). Each of these metrics is binary coded $\{0, 1\}$ and indicates whether the scoring of one of the sleep stages or the arousal is unreliable from PSG when coded with a value of 1. The first sub-metric, MESA quality index, is the sum of these five metrics, and hence ranges from 0 (best) to 5 (worst) points. In addition, we use the electrocardiogram (ECG) signal quality (quecg5), which indicates the

Table 1 Frequency distribution of sleep stages for the MESA and VSL data set, in %

	Wake	NREM	REM	Wake	Light	Deep	REM
MESA	19.2	66.0	14.8	19.2	57.7	8.2	14.8
VSL	11.9	72.3	15.8	11.9	47.4	24.8	15.8

proportion of sleep time where the quality of the ECG signal was good. This metric ranges from 0% (worst) to 100% (best). Finally, a binary coded $\{0, 1\}$ variable (slewake5) indicates whether the quality of the EEG signal only allowed a distinction between *wake* and *sleep*. For the final selection of test nights, we excluded all nights with a MESA quality index above 3, an ECG signal quality below 50% or cases where only *wake* and *sleep* could be distinguished. Consequently, 1826 nights were finally considered for evaluating the models. Due to the nature of the data collection, further pre-processing of the data was necessary. The PSG and the wrist worn actigraphy were attached to the subjects in a clinic. However, the night of sleep took place in the subjects' private homes. Therefore, we trimmed the data using activity counting so that only the data that were actually recorded in the subjects' beds were available. This sample was set up by 53.8% female and 46.2% male participants with an age from 54 to 94 years (mean: 69.1, sd: 8.9).

Self-recorded PSG data (VSL data): Apart from the MESA data, PSG data recorded at the Laboratory for Sleep and Consciousness Research of the Centre for Cognitive Neuroscience (University of Salzburg, Austria) within the VSL project were used as an external test data set. The VSL project aims to validate, through an integrated proof-of-concept prototype, how in situ sleep behaviour detection can be described, measured and assessed using low-cost sensor technologies and methods. All VSL data were recorded with a BrainAmp Standard (Brain Products GmbH 2022) at a sampling rate of 1000 Hz. EEG data were acquired using eight Ag/AgCl electrodes (F3/4, C3/4, O1/2 as well as A1 and A2 for later rereferencing) attached to participants' scalps according to the 10/20 standard system (Jasper 1958). In addition, as requested by the AASM sleep scoring manual (Iber et al. 2007), two electrooculograms and one (bipolar) submental electromyogram were recorded. For heart-rate estimations, a lead II ECG was applied. PSG recordings were automatically classified with Somnolyzer 24x7 (Koninklijke Philips 2022) according to the sleep scoring guidelines of the AASM (Iber et al. 2007) with a post hoc visual quality control by a human expert scorer.[2] R-Peak detection was performed using the open-source Matlab (MATLAB 2020) implementation of the PhysioNet Cardiovascular Signal Toolbox (Vest et al. 2018; Goldberger et al. 2000). Before R-Peak detection, ECG signal was bandpass filtered between 0.1 and 112.5 Hz and downsampled to 250 Hz. In total, our in-house test data set consisted of 46 full-night PSG recordings from 27 different subjects. More specifically, the sample included 18 (66.7%) women and nine (33.3%) men with an age ranging from 20

[2] It should be noted that these sleep stages are based on EEG data and not IBI intervals like the results in this manuscript.

Table 2 Features included in the sleep stage classifier models. All abbreviations and the formulas of the features are described in detail in Martínez et al. (2017)

Type	Features
Respiratory	Mean, sd
Time	Normalised time
Time domain	SDNN, pNN50, SDSD, rMSSD, IRRR, MADRR, TINN, HRV, For IBI and niHR respectively: perc05, perc25, perc50, perc75, perc95, max, min, range, sd, IQR, mad, mean
Frequency domain	For spectral density: perc05, perc25, perc50, perc75, perc95, max, min, range, sd, IQR, mad, mean

to 69 years (mean: 30.6, sd: 12). Thus, our sample was younger on average and included more women than the MESA data set.

To better interpret the following results, it is necessary to mention that the frequency distribution of sleep stages is different for the MESA and the VSL data set. Table 1 shows that MESA has a higher proportion of *wake* (19.2%) than the VSL data set (11.9%). Furthermore, the distribution differs for *deep sleep* for the four sleep stages. Here, we find that *deep sleep* has a much higher proportion with 24.8% in the VSL compared to 8.2% in the MESA data set which was older on average and included more men than the VSL data. This can lead to problems when training the algorithm with the MESA data set. The uneven distribution of sleep stages may lead to that the algorithm trained with the MESA data set tends to classify the waking phase of the VLS data set better, but will have problems correctly classifying deep sleep in the VSL data set. In addition, the algorithm is trained more based on the sleep patterns of older subjects with prior cardiovascular health issues. Since they have different sleep patterns compared to the sample of the VSL data set, there may be challenges in the classification. Furthermore, the *NREM sleep* stage is divided into *light* and *deep sleep* to classify the four sleep stages. *Deep sleep* is only represented here with a share of 8.2%, which makes it difficult for the algorithms to learn its patterns and leads to poorer performance in the classification task for four compared to three sleep stages.

2.2 Feature engineering

By aggregating data from rolling, 270 seconds windows for each 30 second epoch from the annotated R-Peak and respiratory data of the PSG 46 features were calculated. The HRV features were obtained with the R package RHRV (Rodriguez-Linares 2020), are either based on time-domain or frequency-domain features and measure the variability of the heart rate (Martínez et al. 2017). All HRV features listed in Table 2 and their formulas are described in detail in Martínez et al. (2017). In addition, standard statistics, such as mean, standard deviation or percentiles, were calculated for each window. The respiratory features were obtained from an airflow thermistor. Furthermore, a time feature represents the normalised time spent by a person between going to bed (0) and getting up (1). Therefore, 47 features that are used in the random forest model have been calculated in total. All features are listed in Table 2. Table 9 in the Appendix displays the descriptive statistics (mean and standard deviation) for each feature.

2.3 Random forests

Random forests (Breiman 2001) are nowadays one of the most popular classification and regression methods for practitioners, because they exhibit good performance on a wide variety of tasks (Biau and Scornet 2016; Efron and Hastie 2016). They are an ensemble method that combines diverse, tree-shaped classifiers to improve over their individual performance. Diversity is achieved by injecting randomness into the induction process, specifically by bootstrapping the data and sampling from the available attributes. The random forest was grown with the ranger package (Wright and Ziegler 2017) in the statistical software R (R Core Team 2020). For hyperparameter tuning, we used the caret (Kuhn 2020) package in R. To find the optimal number of trees drawn with standard bootstrapping samples (`num.trees`), we gave three different options: 100, 300 and 500. For the search for the number of variables to be randomly selected for each split (`mtry`), we specified a tune length of 10 and tried the following values for `mtry`: 2, 7, 12, 17, 22, 27, 32, 37, 42, 47. For both classification tasks, the best model parameters found were a `mtry` of 22. For `num.trees`; however, we find no significant changes in performance when using 100, 300 or 500, as only the third decimal place of the kappa or accuracy value changes in each case. However, as a `num.trees` of 500 gave the best results, we decided to use this figure for the evaluation.

2.4 Deep neural networks

Artificial neural networks became extremely popular (again) after a CNN won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012 (Krizhevsky et al. 2012). This led to what is sometimes called the 3rd golden age of neural network research, in which deep multi-layer perceptrons, that is, feed-forward neural networks with many hidden layers, are used for complex classification and prediction tasks. The advantages of neural networks are that no expert knowledge is required, that they can be trained “end-to-end" and that these models can be better generalised because they are less susceptible to interference. For time series analysis, as we face it here, also LSTM models, as introduced in the seminal paper of Hochreiter and Schmidhuber (1997) and later extended from Gers et al. (2000), showed superior performance to classical methods. For an extensive overview on deep learning approaches, see Goodfellow et al. (2015).

In contrast to the random forest, the engineered features were not used for the two neural networks: (1) CNN-LSTM and (2) seq2seq. For these two models, the time series of IBIs and respiration rates were resampled to 2 Hz and 0.5 Hz, respectively, and directly included. To extract local features, the input signals were at first fed into convolutional layers in both models. The following part of the models involved learning longer range patterns that provided the necessary information for the sleep stage classification. This part differed between the two neural networks in the amount of information used to classify each epoch and in this way, longer term dependencies were learned.

The first model (*CNN-LSTM*) received single 270 seconds windows as input, for which the sleep stage of the central 30 second epoch should be classified, which is similar to the random forest model. It consisted of a feature extraction part that involved three layers of residual CNN blocks for the IBI time series and two for the sequence of respiration rates. A single residual block consisted of two convolutional layers each preceded by a leaky rectified linear unit activation function and batch normalisation as well as a residual connection at the end. The kernel sizes of the convolutional layers increased with each residual block (5, 10 and 15 for the IBI sequence, 5 and 10 for the sequence of respiration rates). This was followed by two layers of bidirectional recurrent units, each containing 64 LSTM cells with tanh (Goodfellow et al. 2015) activation functions. The results were fed into a dense layer with a softmax activation that performed the final classification. To avoid overfitting, a dropout rate of 25% was used in the recurrent units. For optimisation, Adam (Kingma and Ba 2014) with a learning rate of 0.0001 was used. The model was trained over 50 epochs with a batch size of 64. However, the training process was stopped once the loss on the validation data did not decrease for eight consecutive epochs.

The second model (*seq2seq*) was a sequence to sequence model that received the entire IBI and respiration rate time series of each individual night and outputs a sequence of sleep stage predictions, one for each 30 second epoch. It is essentially the model proposed by Sridhar et al. (2020) with a few adjustments: First, the signal of the respiration rate was included in the model. The local features of this signal were extracted similarly to the local IBI features and the results were concatenated before the dilated convolutional blocks. “Local" in this context refers to a 128 s segment of the respective signal as both signals were represented by a sequence of 1200 overlapping batches, each covering 128 seconds and centered around the 30 second epoch for which the sleep stage should be predicted. Second, L1 regularisation of the weights was omitted, resulting in an increase in the model’s performance when training with the MESA data set. Finally, similar to the first network, this model was also subject to early stopping.

Categorical cross-entropy was used as loss function in both of the above described models. To consider the imbalanced distribution of classes in the data set, we followed Sun et al. (2020) and implemented a weighted loss calculation, where the weights corresponded to the inverse class frequency in the training data.

2.5 Evaluation

To answer the first research question and evaluate the generalisation performance of random forests and neural networks, two steps were performed. In the first step, we applied a fivefold cross-validation to the MESA data set, which gives us an overview of the general model performance of the respective model. In a second step, we used the trained algorithm of the respective model from the first step and applied it to the self-recorded VSL data.

Thus, we can interpret the generalisation performance of our models when we compare the results between the first and the second step, the two classification tasks and across the different models. The first research question is, thus, answered by comparing the classification performance between the results with the MESA test data and the VSL data for each method separately. To answer the second research question, the classification results of the three methods are compared when applied to the VSL data.

For the validation of all classification approaches, accuracy, Cohen's Kappa, weighted F1-scores and confusion matrices will be provided. Cohen's Kappa (Cohen 1960) provides a measure of accuracy that excludes random accuracy. Its value ranges from −1 (worst possible classification) to +1 (perfect classification). A Cohen's Kappa of 0 would mean that the results correspond to a random classification. The F1-score is a metric that calculates the harmonic mean between recall and precision, where recall measures the true positive predicted classes in relation to all positive cases and the precision measures the true positive predicted classes in relation to all positive predictions (Powers 2011). For the model evaluation, we applied a fivefold cross-validation and calculated the average of the corresponding performance metric across all five splits for each of the three machine learning approaches and reported the average and the standard deviation.

To show whether the classification results of the respective methods are similar or not, and to facilitate the comparison of the classification results, we use the Adjusted Rand Index (ARI) (Rand 1971; Hubert and Arabie 1985) as a summary measure. With the ARI, we compare the predicted sleep stages between each of the three applied models, respectively. The ARI ranges between zero (completely distinct results) and one (completely similar results).

3 Results

Table 3 shows the accuracies, Cohen's Kappa and weighted F1-score for all three classification approaches and both classification tasks (three sleep stages, as well as four sleep stages). For three classes, the accuracies ranged from 68.3% for the random forest to 82.1% for the seq2seq model when using the MESA data for validation. Furthermore, for four classes, the results ranged from 59.9% for the random forest to 74.2% for the seq2seq model for the MESA data. For the VSL data, the random forest showed classification accuracies of 73.3% for three and 49.4% for four classes. The seq2seq model achieved classification accuracies of 81.2% for three and 64.4% for four classes for the VSL data. Table 3 shows that all three classification methods are generalisable for three sleep stages from the MESA to the VSL data, as the accuracy and Cohen's Kappa values were quite similar for

Table 3 Mean values of fivefold cross-validation for classification accuracy, Cohen's kappa and F1 scores of sleep stage classifiers

		MESA data			VSL data		
		Accuracy in %	Cohen's Kappa	Weighted F1-score in %	Accuracy in %	Cohen's Kappa	Weighted F1-score in %
3 classes	RF	68.3 (0.6)	0.237 (0.009)	63.5 (0.6)	73.3 (0.1)	0.189 (0.004)	63.6 (0.3)
	CNN-LSTM	79.5 (0.4)	0.537 (0.014)	77.7 (0.5)	81.0 (0.4)	0.506 (0.022)	79.4 (0.6)
	seq2seq	82.1 (0.9)	0.641 (0.012)	81.7 (0.5)	81.2 (0.5)	0.571 (0.02)	81.2 (0.6)
4 classes	RF	59.9 (1.07)	0.194 (0.01)	53.8 (1.2)	49.4 (0.2)	0.096 (0.004)	45.8 (0.8)
	CNN-LSTM	71.6 (0.5)	0.45 (0.014)	67.1 (1.2)	56.4 (0.5)	0.267 (0.025)	47.3 (0.8)
	seq2seq	74.2 (0.8)	0.553 (0.014)	73.0 (0.9)	64.4 (0.6)	0.444 (0.023)	61.9 (0.7)

Standard deviation in parentheses

Table 4 Confusion matrices of sleep stage classifiers based on MESA test data, three sleep stages

		RF predicted			NN predicted			s2s predicted		
		Wake	NREM	REM	Wake	NREM	REM	Wake	NREM	REM
actual	Wake	23.1	68.5	8.4	43.9	52.2	3.9	62.6	31.9	5.5
	NREM	4.5	92.4	3.1	2.5	95.5	2.0	5.5	90.7	3.8
	REM	10.9	69.6	19.5	4.1	41.2	54.6	6.7	22.1	71.2

Mean values of fivefold cross-validation for classification (RF: random forest, NN: CNN-LSTM, s2s.: seq2seq model)

Table 5 Confusion matrices of the sleep stage classifiers based on the MESA test data set, four sleep stages

		RF predicted				NN predicted				s2s predicted			
		Wake	Light	Deep	REM	Wake	Light	Deep	REM	Wake	Light	Deep	REM
actual	Wake	24.6	66.3	0.2	8.9	47.8	47.9	0.0	4.3	62.4	31.2	0.3	6.1
	Light	5.5	89.6	1.1	3.7	3.5	94.2	0.1	2.3	6.2	86.0	3.3	4.5
	Deep	2.8	90.0	6.3	0.9	1.5	97.7	0.4	0.7	1.8	69.3	27.3	1.6
	REM	11.6	67.9	0.1	20.4	5.0	39.4	0.0	55.6	6.1	21.0	0.2	72.6

Mean values of fivefold cross-validation for classification (RF: random forest, NN: CNN-LSTM, s2s: seq2seq model)

Table 6 Confusion matrices of sleep stage classifiers based on the VSL test data, three sleep stages

		RF predicted			NN predicted			s2s predicted		
		Wake	NREM	REM	Wake	NREM	REM	Wake	NREM	REM
actual	Wake	20.4	76.3	3.3	53.6	42.9	3.6	53.7	38.8	7.5
	NREM	3.0	95.8	1.2	3.5	94.6	1.9	5.4	88.7	5.8
	REM	9.2	80.8	9.9	5.2	54.6	40.2	8.0	24.1	67.9

Mean values of fivefold cross-validation for classification (RF: random forest, NN: CNN-LSTM, s2s: seq2seq model)

Table 7 Confusion matrices of sleep stage classifiers based on the VSL test data, four sleep stages. Mean values of fivefold cross-validation for classification (RF: random forest, NN: CNN-LSTM, s2s: seq2seq model)

		RF predicted				NN predicted				s2s predicted			
		Wake	Light	Deep	REM	Wake	Light	Deep	REM	Wake	Light	Deep	REM
actual	Wake	22.4	73.9	0.0	3.7	61.3	35.0	0.0	3.7	49.2	40.7	0.5	9.6
	Light	4.2	93.9	0.3	1.7	6.7	90.7	0.0	2.7	5.8	83.4	1.2	9.6
	Deep	1.7	95.6	2.2	0.5	2.6	96.8	0.0	0.6	1.6	68.5	28.4	1.5
	REM	9.5	80.0	0.0	10.4	8.0	53.4	0.0	38.2	3.8	20.8	0.0	75.4

Table 8 Adjusted Rand Index based on predicted values of VSL or MESA test data and three or four sleep stages (RF: random forest, NN: CNN-LSTM, s2s: seq2seq model)

Test data set	Sleep stages	RF vs. NN	RF vs. s2s	NN vs. s2s
MESA	Three	0.71	0.68	0.77
MESA	Four	0.69	0.58	0.68
VSL	Three	0.74	0.61	0.69
VSL	Four	0.75	0.43	0.53

both data sets. However, when classifying four sleep stages, all three classification approaches perform worse for VSL data compared to the MESA test data set.

Tables 4 and 5 show the confusion matrices when using the MESA data set for validation. We found that the random forest showed the best results for *NREM sleep* with a percentage of 92.4% of correctly classified epochs. For *NREM sleep* the CNN-LSTM showed classification accuracies of 95.5% and the seq2seq model of 90.7%. For four sleep stages the random forest performed best for *light sleep*, with a proportion of 89.6% correctly classified epochs. For four sleep stages the CNN-LSTM and the seq2seq model showed the best classification performance for *light sleep* (CNN-LSTM: 94.2%; seq2seq model: 86.0%).

Tables 6 and 7 show the results using the VSL data for validation. For three sleep stages, *NREM sleep* (95.8%) and *wake* (20.4%) showed similar results compared to Table 4, when the MESA data were used for validation for the random forest. For four sleep stages, the classification performance for *REM sleep* dropped to 10.4%

and for *deep sleep* to 2.2%, whereas the other two sleep stages showed similar or even better results regardless of whether the MESA or VSL test data were used to validate the random forest. Also for the CNN-LSTM and the seq2seq model, *NREM* was best classified in the three stage classification task (94.6% and 88.7%, respectively), while *light sleep* was best classified in the four stage classification task with 90.7% (CNN-LSTM) and 83.4% (seq2seq).

Table 8 shows the ARI that ranged between a value of 0.43 (random forest vs. seq2seq) for the VSL test data set and the four-level classification task and 0.77 (CNN-LSTM vs. seq2seq) for the MESA test data set and the three-level classification task. In general, we find that in most model comparisons, the ARI is higher in the three-stage classification task compared to the four-stage classification task especially when comparing the other models to the seq2seq. This is likely due to the fact that the other two models label almost no epochs as *deep sleep* whereas the seq2seq model has at least a decent ability to detect *deep sleep*.

4 Discussion

When analysing the classification performance of single classes, in the three-stage classification task, *NREM* was classified best, while *wake* was classified worst. Furthermore, in the four-stage classification task, *light sleep* was classified best, while *deep sleep* was classified worst. These findings applied to all three classification approaches.

As of today, human experts represent the gold standard in the classification of sleep. However, it is well known that the inter-rater agreement between human experts is far from perfect. More specifically, human experts agree for five classes in only 82.6% (Danker-Hopfe et al. 2004) or for four classes in 88% (Sridhar et al. 2020; Rosenberg and Van Hout 2013) when staging sleep based on PSG data. Hence, the interpretation of the performance of automatic sleep classification procedures as described in this paper should be based on these imperfect human expert benchmarks. With regard to the first research question, we found that each of the three models showed similar classification performance for three sleep stages when either MESA or self-recorded VSL data were used for validation. While in the three stage classification task, the weighted F1-score equals 63.5% and 63.6% for the MESA as well as VSL data set, it drops from 53.8% (MESA) to 45.8% (VSL) for the four class classification task for the random forest model. Also for the CNN-LSTM and seq2seq model, the weighted F1-score is similar for the MESA and VSL validation data for three classes, while it drops for four classes. However, when reducing the granularity by dividing the *NREM* stage into *light* and *deep sleep*, the classification performance was slightly worse when VSL data were used for validation. This allows us to support the first research question for the three-stage classification task and establish that all three models are generalisable to our self-recorded VSL data.

On the other hand, the research question for four classes can only be supported conditionally, as the results for the VSL test data were slightly worse and did not correspond to the results when using the MESA data for validation. In particular, the *deep sleep* stage class was predicted worse with the VSL test data. This sleep stage was relatively less frequent in the MESA data than in the VSL test data and was predicted worse with all three classification models in the VSL test data set. Regarding the second research question, we found that the three models differed in their classification performance. We found that the seq2seq model showed the best performance, followed by the CNN-LSTM model and the random forest, which showed the poorest performance. Thus, we can state that the seq2seq model generalises best to our data and has the best classification performance, when using self-recorded VSL data. The results also indicate that information about the broader temporal context of the respective sleep epoch is beneficial for the overall classification performance, especially as far as the classification of *deep sleep* phases is concerned. This becomes obvious when comparing the classification performance of the seq2seq model with the other two for the four-stage classification task. While the seq2seq model receives the IBI time series of the entire night as input and can, therefore, base its predictions for every single epoch on a very broad temporal context, the other two models feature a scope limited to 270 s. Also, the seq2seq model appears to be the only one showing at least a decent performance when classifying *deep sleep*, while the other two models seem to be almost unable to detect *deep sleep* epochs.

In addition, the ARI shows us the degree of agreement between the models and how their classification results differ between them. Comparing the models with the same test data set and classification level, we find the lowest ARIs between the random forest and seq2seq models, while there is no clear picture for the highest values. For the VSL data, we found the highest agreement of the models' result between the random forest and the CNN-LSTM for the three, as well the four classification task. While for the MESA data, the highest agreement for the three classification task can be seen between CNN-LSTM and seq2seq model, the agreement between the random forest and CNN-LSTM and between CNN-LSTM and seq2seq is very similar for the four sleep stage classification results.

Nevertheless, even the performance of the seq2seq model is rather low in comparison to other deep learning approaches that were using IBI data and deep learning models such as the one developed in the work of Sridhar et al. Sridhar et al. (2020), even though the architecture of the seq2seq model is heavily based on this work. The reason for this is twofold. First, the MESA data set used for training is comparatively small to data sets used in other works especially considering the relatively high number of parameters of the seq2seq model (e.g. Sridhar et al. (2020) used a combined set of the MESA and SHHS data set resulting in a total of 10, 332 nights). Second, the main aim of this work is not to maximise the models' performance on a specific data set but to develop a model that generalises well to a data set that differs from the training set in several aspects. Since the test data set comprised participants of a much younger age (the average age is 69.6 for the MESA data set and 32.7 for the VSL data set) with no prior cardiovascular condition as is the case for the MESA data set, this indeed presents a challenge to the generalisability of the model. Future work within the VSL project will concentrate on the four sleep stage classification task and further classification models, e.g. transfer learning (Radha et al. 2021)-based approaches and models containing additional data sets will be developed to improve the four stage classification models.

For the robustness of the models, we applied a fivefold cross-validation. Future work could further test the robustness of the developed models using different cohorts of subjects (e.g. subjects with different disease patterns) to investigate whether the models are still generalisable to these different subjects' characteristics.

5 Conclusion

This work demonstrated the generalisability of sleep stage classification and its challenges from three machine learning approaches using a new self-recorded data set. The results showed that the seq2seq model is best generalisable to the self-recorded sleep data set. In addition, we find that for the three-stage classification task—*wake, NREM sleep, REM sleep*—all three machine learning approaches are suitable to classify sleep stages of a new data set. However, classifying the four sleep stages—*wake, light sleep, deep sleep, REM sleep*—showed that the model trained with the MESA data set is not well generalisable to the VSL data, as especially *deep sleep* was predicted worse with the VSL data compared to the MESA data set. This implies that more data sets or other classification approaches, e.g. transfer-learning approaches, are needed to train models which are also generalisable for four sleep stages.

Appendix

Table 9 displays the descriptive statistics of all 47 features used in the random forest model. All abbreviations and the formulas of the features are described in detail in Martínez et al. (2017).

Table 9 Descriptive statistics of all 47 features

Variable name	Mean	sd
Respiratory		
Resp_mean	15.65	1.77
Resp_sd	0.85	0.45
Time		
Epoch_norm	100	14.99
Time domain		
SDNN	57.02	73.71
pNN50	13.25	18.45
SDSD	44.85	78.13
rMSSD	44.77	78.08
IRRR	64.32	54.61
MADRR	21.94	28.55
TINN	156.85	72.00
HRVi	10.04	4.61
RR_perc_05	864.54	141.67
RR_perc_25	923.31	144.87
RR_perc_75	987.63	156.65
RR_perc_95	1032.80	170.17
RR_max	1141.16	747.44
RR_min	778.11	598.20
RR_range	363.05	953.25
RR_sd	57.02	73.71
RR_med	956.72	149.83
RR_IQR	64.32	54.61
RR_mad	45.10	34.94
RR_mean	954.24	147.09
niHR_perc_05	59.65	9.75
niHR_perc_25	62.28	9.95
niHR_perc_75	66.62	10.77
niHR_perc_95	71.32	12.19
niHR_max	77.79	16.33
niHR_min	55.39	11.46
niHR_range	22.40	17.07
niHR_sd	3.79	2.83
niHR_med	64.27	10.25
niHR_IQR	4.33	3.95
niHR_mad	3.01	2.52
niHR_mean	3.01	2.52
Frequency domain		
spec_max	120599.36	339935.22
spec_min	3056.18	91439.46
spec_sd	13744.89	36299.43
spec_mean	8989.44	139988.68

Table 9 (continued)

Variable name	Mean	sd
spec_med	5724.49	138777.24
spec_q25	4614.45	126856.12
spec_q75	7936.15	151992.42
spec_q95	22719.94	176052.81
spec_q05	3777.93	112672.20
spec_mad	1921.70	25092.60
spec_IQR	3321.70	33817.98
spec_range	117543.18	299256.55

sd standard deviation

Acknowledgements The 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-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).

Funding The authors gratefully acknowledge the financial support from Land Salzburg within the WISS 2025 project Virtual Sleep Lab (VSL-Lab) (20102-F2002176-FüR) and the WISS 2025 project IDA-Lab (20102-F1901166-KZP and 20204-WISS/225/197-2019).

Data availability For model training, we used the Multi-Ethnic Study of Atherosclerosis (MESA) data set. This data set can be requested under the following link: https://sleepdata.org/data/requests/mesa/start. The self-recorded PSG data (Virtual Sleep Lab (VSL) data) are available from the corresponding author on reasonable request.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Biau G, Scornet E (2016) A random forest guided tour. Test 25:197–227

Bild DE (2002) Multi-ethnic study of atherosclerosis: objectives and design. Am J Epidemiol 156(9):871–881

Brain Products GmbH (2022) Brainamp standard. https://www.brainproducts.com/productdetails.php?id=1. Accessed 13 Apr 2023

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Chattu VK, Manzar M, Kumary S, Burman D, Spence DW, Pandi-Perumal SR et al (2019) The global problem of insufficient sleep and its serious public health implications. Healthcare 7(1):1–16

Chen X, Wang R, Zee P, Lutsey PL, Javaheri S, Alcántara C, Jackson CL, Williams MA, Redline S (2015) Racial/ethnic differences in sleep disturbances: the multi-ethnic study of atherosclerosis (MESA). Sleep 38(6):877–888

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Measurement 20(1):37–46

Danker-Hopfe H, Kunz D, Gruber G, Klösch G, Lorenzo JL, Himanen SL, Kemp B, Penzel T, Röschke J, Dorn H et al (2004) Interrater reliability between scorers from eight european sleep laboratories in subjects with different sleep disorders. J Sleep Res 13(1):63–69

Efron B, Hastie T (2016) Computer age statistical inference. Cambridge University Press, New York

Faust O, Razaghi H, Barika R, Ciaccio EJ, Acharya UR (2019) A review of automated sleep stage scoring based on physiological signals for the new millennia. Comput Methods Programs Biomed 176:81–91

Gers FA, Schmidhuber J, Cummins F (2000) Learning to forget: continual prediction with LSTM. Neural Comput 12:2451–2471

Ghassemi MM, Moody BE, Lehman LWH, Song C, Li Q, Sun H, Mark RG, Westover MB, Clifford GD (2018) You snooze, you win: the physionet/computing in cardiology challenge 2018. In: 2018 Computing in Cardiology Conference (CinC), IEEE, vol 45, pp 1–4

Goldberger A, Amaral L, Glass L, Hausdorff J, Ivanov PC, Mark R, Mietus JE, Moody GB, Peng CK, Stanley HE (2000) Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation [Online] 101(23):e215–e220

Goodfellow I, Bengio Y, Courville A (2015) Deep Learning. MIT Press, Cambridge

Guillot A, Thorey V (2021) Robustsleepnet: Transfer learning for automated sleep staging at scale. arXiv preprint arXiv:2101.02452

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9:1735–1780

Hubert L, Arabie P (1985) Comparing partitions. J Classif 2(1):193–218

Iber C, Ancoli-Israel S, Chesson A, Quan S (2007) The aasm manual for the scoring of sleep and associated events: rules, terminology and technical specifications. American Academy of Sleep Medicine, Westchester

Jasper H (1958) The ten-twenty electrode system of the international federation. Electroencephalogr Clin Neurophysiol 10:371–375

Jiang D, Lu Y, Ma Y, Wang Y (2019) Robust sleep stage classification with single-channel EEG signals using multimodal decomposition and HMM-based refinement. Expert Syst Appl 121:188–203

Kerkhof GA (2017) Epidemiology of sleep and sleep disorders in the Netherlands. Sleep Med 30:229–239

Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980

Koninklijke Philips NV (2022) Somnolyzer 24x7. https://www.philips.com.hk/healthcare/product/HC1076888/sleep-diagnostic-somnolyzer-24x7-scoring-solution-sleep-scoring-software. Accessed 13 Apr 2023

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Commun ACM 60:84–90

Kuhn M (2020) caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret. Accessed 13 Apr 2023

Loh HW, Ooi CP, Vicnesh J, Oh SL, Faust O, Gertych A, Acharya UR (2020) Automated detection of sleep stages using deep learning techniques: a systematic review of the last decade (2010–2020). Appl Sci 10(24):8963

Martínez CAG, Quintana AO, Vila XA, Touriño MJL, Rodríguez-Liñares L, Presedo JMR, Penín AJM (2017) Heart rate variability analysis with the R package RHRV. Springer, Cham, Switzerland

MATLAB (2020) MATLAB version 9.9.0.1538559 (R2020b) Update 3. The Mathworks, Inc., Natick

Olesen AN, Jennum PJ, Mignot E, Sorensen HBD (2020) Automatic sleep stage classification with deep residual networks in a mixed-cohort setting. Sleep 44(1):zsaa161

Powers DM (2011) Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. J Mach Learning Technol 2(1):37–63

Quan SF, Howard BV, Iber C, Kiley JP, Nieto FJ, O'Connor GT, Rapoport DM, Redline S, Robbins J, Samet JM et al (1997) The sleep heart health study: design, rationale, and methods. Sleep 20(12):1077–1085

R Core Team (2020) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. Accessed 13 Apr 2023

Radha M, Fonseca P, Moreau A, Ross M, Cerny A, Anderer P, Long X, Aarts RM (2019) Sleep stage classification from heart-rate variability using long short-term memory neural networks. Sci Rep 9(1):14149

Radha M, Fonseca P, Moreau A, Ross M, Cerny A, Anderer P, Long X, Aarts RM (2021) A deep transfer learning approach for wearable sleep stage classification with photoplethysmography. NPJ Digital Med 4(1):1–11

Rand WM (1971) Objective criteria for the evaluation of clustering methods. J Am Stat Assoc 66(336):846–850

Rodriguez-Linares L, Vila X, Lado MJ, Mendez A, Otero A, Garcia CA (2020) RHRV: Heart Rate Variability Analysis of ECG Data. R package version 4.2.6. https://CRAN.R-project.org/package=RHRV. Accessed 13 Apr 2023

Rosenberg RS, Van Hout S (2013) The American academy of sleep medicine inter-scorer reliability program: sleep stage scoring. J Clin Sleep Med 9(1):81–87

Sridhar N, Shoeb A, Stephens P, Kharbouch A, Shimol DB, Burkart J, Ghoreyshi A, Myers L (2020) Deep learning for automated sleep staging using instantaneous heart rate. npj Digital Med 3(1):106

Sun H, Ganglberger W, Panneerselvam E, Leone MJ, Quadri SA, Goparaju B, Tesh RA, Akeju O, Thomas RJ, Westover MB (2020) Sleep staging from electrocardiography and respiration with deep learning. Sleep 43(7):zsz306

Vest AN, Poian GD, Li Q, Chengyu Liu, Nemati S, Shah A, Clifford GD (2018) Cliffordlab/physionet-cardiovascular-signal-toolbox: Physionet-cardiovascular-signal-toolbox 1.0. https://doi.org/10.5281/ZENODO.1243112. Accessed 13 Apr 2023

Wright MN, Ziegler A (2017) Ranger: a fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw 77(1):1–17

Zhai B, Perez-Pozuelo I, Clifton EAD, Palotti J, Guan Y (2020) Making sense of sleep. Proc ACM Interact Mobile Wearable Ubiquitous Technol 4(2):1–33

Zhang GQ, Cui L, Mueller R, Tao S, Kim M, Rueschman M, Mariani S, Mobley D, Redline S (2018) The national sleep research resource: towards a sleep data commons. J Am Med Inform Assoc 25(10):1351–1358

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.