Sleep Classification Using Consumer Sleep Technologies and AI

A review of the current landscape

Djanian, Shagen; Bruun, Anders; Nielsen, Thomas Dyhre

Published in:
Sleep Medicine

DOI (link to publication from Publisher):
10.1016/j.sleep.2022.09.004

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Sleep Medicine 100 (2022) 390–403

Contents lists available at ScienceDirect

Sleep Medicine

journal homepage: www.elsevier.com/locate/sleep

Sleep classification using Consumer Sleep Technologies and AI: A review of the current landscape

Shagen Djanian*, Anders Bruun, Thomas Dyhre Nielsen

Aalborg University, Department of Computer Science, Denmark

ARTICLE INFO

Article history:
Received 21 June 2022
Accepted 5 September 2022
Available online 22 September 2022

Keywords:
Sleep
Artificial intelligence
Interaction
Intervention

ABSTRACT

Classifying sleep stages in real-time represents considerable potential, for instance in enabling interactive noise masking in noisy environments when persons are in a state of light sleep or to support clinical staff in analyzing sleep patterns etc. However, the current gold standard for classifying sleep stages, Polysomnography (PSG), is too cumbersome to apply outside controlled hospital settings and requires manual as well as highly specialized knowledge to classify sleep stages. Using data from Consumer Sleep Technologies (CSTs) to inform machine learning algorithms represent a promising opportunity for automating the process of classifying sleep stages, also in settings outside the confinements of clinical expert settings. This study reviews 27 papers that use CSTs in combination with Artificial Intelligence (AI) models to classify sleep stages. AI models and their performance are described and compared to synthesize current state of the art in sleep stage classification with CSTs. Furthermore, gaps in the current approaches are shown and how these AI models could be improved in the near-future. Lastly, the challenges of designing interactions for users that are asleep are highlighted pointing towards avenues of more interactive sleep interventions based on AI-infused CSTs solutions.

1. Introduction

Getting a good nights sleep is critical for our well being. It is well-known that disruptions to our sleep lead to various negative consequences for our health, quality of life, and work productivity. Short-term consequences for our health are increased stress level, somatic problems, emotional distress and mood disorders [1]. In terms of productivity, studies have shown that short-term consequences of sleep disturbance lead to cognitive and performance deficits, cf [1, 2]. The risk of work related injuries are also shown to increase by a factor of 1.62 due to sleep issues [3]. While there is a large body of knowledge on how to reduce the negative effects of lifestyle, psychological and medical conditions on sleep, they mostly focus on sleep hygiene [4] based on approaches of preventive measures. These are mainly related to lifestyle changes. An example of an area where preventive measurements are limited, are sleeping in noisy environments. There are rather few studies dealing with the challenges of reducing the effect of noise disruptions of our environment [5]. alre Another solution is using loudspeakers to mask environmental sounds using white noise. This has shown to have positive short-term effect on improving sleep quality [6] while recent studies suggest that prolonged exposure to noise from sound machines can induce auditory perceptual problems [7]. Instead of a static solution, where e.g. a white noise machine is either turned on the whole night or turned off with a timer, this calls for a more interactive solution. By monitoring the sleep stages of a user it could become possible to intervene regarding the various disturbing factors that can arise in the night, in real time, and particularly during light sleep.

Monitoring and measurement of sleep is a vast area of research where many different approaches have been utilized. The study by Ref. [8] presents a review of methods to assess sleep quality and a ranking of existing methods including questionnaires and diaries [8], contactless devices [9], contact devices [10] and Polysomnography (PSG) [11]. The gold standard for sleep studies involves PSG which is a device consisting of multiple sensors; Electroencephalography (EEG) for measuring brain activity, Electrooculography (EOG) for eye movements, Electromyography (EMG) for muscle activity or skeletal muscle activation, and Electrocardiography (ECG) for heart rhythm [12]. These recordings are then manually scored by trained experts to classify the sleep. They typically classify sleep according to the American Academy of Sleep

* Corresponding author.
E-mail address: shagendj@cs.aau.dk (S. Djanian).

https://doi.org/10.1016/j.sleep.2022.09.004

Medicine (AASM) guidelines [13] which involve 5 stages:

- Stage W (Wakefulness)
- Stage N1 (NREM 1)
- Stage N2 (NREM 2)
- Stage N3 (NREM 3)
- Stage R (REM)

There are a multitude of reasons why a PSG is not always feasible with some major drawbacks being that they are resource expensive and obtrusive [14, 15]. A promising approach to solve these problems are Consumer Sleep Technologies (CSTs) [16]. They are both cheaper and less intrusive than PSG but unfortunately they are not yet suitable for clinical studies [17]. Their sleep stage classification is currently not at the level of PSGs and the products often have a proprietary algorithm that makes it a black box with no access to the raw data. Nonetheless, interest in CSTs has risen in recent years. A promising approach with CSTs are using the physiological data in combination with the advances made in Artificial Intelligence (AI) the past 10 years. These advances open a path to measuring and improving sleep in a new way. Typically it has been most common to assess sleep quality and disruption in the night through indirect measures based on questionnaires after one has finished sleeping [8]. With CSTs a more interactive approach based on Machine Learning (ML) and real-time noise masking may be possible for improving sleep.

To be able to have a more interactive approach towards sleep intervention it is relevant to understand the current landscape of research into AI based CSTs. The review by Ref. [18] covers a broad area of computational sleep research including sleep classification for medical and home use. Likewise [19] is a similar study of Deep Learning (DL) for sleep classification but only covers PSG based studies. We expand previous research efforts by providing 1) an overview of state-of-the-art on models for sleep stage classification based on CST data and 2) a performance analysis of these classification models highlighting avenues for future research and implications towards development of CSTs suitable for home use.

The main contribution of this article is two-fold. Firstly *we provide a summary of CST devices and ML algorithm performance in classifying sleep stages*. Throughout this review we aim to provide a future outlook for researchers outside the area of machine learning based on our knowledge as researchers within the field of computer science. We aim to strike a balance between giving an overview of the fundamental ML techniques and performance metrics for researchers outside this domain of expertise while at the same time emphasizing state-of-the-art. Secondly *we open a discussion on the implications of designing a system where users interacts with the system while they are asleep*.

2. Machine learning

ML is a central part of AI and the two are often used interchangeably. This section provides a general and simplified understanding of ML and is intended for researchers with limited familiarity of ML. It can be seen as a family of mathematical models that optimize their model coefficients by reducing a loss function on a given dataset. This process is referred to as training a model, and the dataset used for this is the training data. A part of the dataset is often withheld from the model during training and used afterwards to test how well the model performs on unseen data. This part of the dataset is called the test data. These models can then be used for tasks such as classification or regression. They have been widely successful in fields such as medicine, speech recognition and computer vision [20]. One of the main reasons that ML has proven successful in the last two decades is that advances in computing power has allowed these models to be trained effectively on massive amounts of data in a reasonable time frame [21]. Another reason is that massive amounts of data have become available like the crowd sourced image dataset of 1.2 million images used in training AlexNet [22]. Traditionally the flow of these models has been to perform feature engineering on the datasets to find features that represent important relationships between predictors in the datasets. This approach is especially powerful when the dataset is low dimensional and limited data is available [20] These features are typically chosen by a combination of previous domain knowledge and trial and error. Examples of this in the domain of sleep stage classification are Heart Rate Variability (HRV) features. HRV has been used by Ref. [23] where they calculated various features like the mean and standard deviation of interbeat intervals, and power of the various frequency bands. Similarly this is done in actiography where features are calculated from an accelerometer signal and fed to a Support Vector Machine (SVM) to predict sleep or awake. This approach has the benefit of using input that has been well studied and builds on an existing foundation. The features are transparent, easily replicable and well understood. The drawback is the limitation of finding the right features to represent the signal which requires a great deal of feature engineering.

DL is a subset of ML that has had massive success in the last decade and are outperforming classical ML models tasks where the data is large scale, noisy and unstructured [20]. DL models are based on neural networks, where a neuron is linked to many other neurons in various configurations. Collectively, their configurations are referred to as the neural networks architecture and they allow the networks to learn highly complex and non-linear relationships between input and output. Earlier neural network architectures are called feed-forward Multilayer Perceptron (MLP) as the input is fed forward through a fully connected network with multiple layers of neurons and backpropagation is performed to optimize each neuron according to a loss function. This is done in multiple iterations called epochs until model performance is stabilized. Architectures like Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) have different configurations of neurons that allow them to learn different kinds of features, like spatial or temporal features. Even though they have a different architecture they still build on the principles from MLP with using backpropagation to optimize neurons. What makes DL different from classical ML models is that while they can be used with features from feature engineering as input they can also use the raw data as input instead. Examples of these can be images or audio files. The model can then learn internal features that it will use to perform classification or regression. A simplified graphical representation of the difference in the process between feature engineering and the DL can bee seen in Fig. 1. The drawback of this is that the model becomes more like a black box where it can be difficult to determine what it actually learns and what it bases its decision making on. This runs the risk of the model learning a relationship between the input and output that might only be applicable to the specific dataset. Explainability and transparency of these DL models is also a field of study that is getting more attention [24]. Nonetheless they have shown great progress and are a promising approach when working with biological signals.

3. Method

A systematic review of the recent application of ML and DL models to automatically classify sleep stage with input from CSTs conducted and is here reported accordingly to the PRISMA statement guidelines [25]. A detailed illustration of the selection process to extract the relevant papers can be seen in Fig. 3.

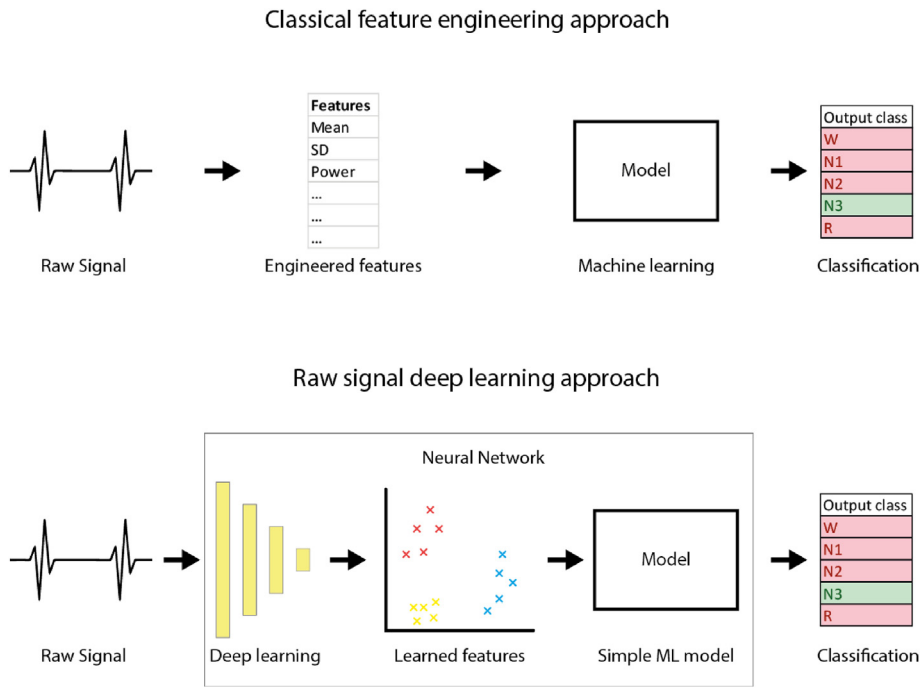

Fig. 1. A simplified process of feature based ML approach and DL that uses the raw signal. The difference between them is whether to base the model on features that are known to represent the signal well or let the DL model derive its own features.

- **Search strategy:** Electronic searches in Scopus were performed. To get a better understanding of the terminology used in sleep stage classification, a naïve search was done. To ensure a high level of inclusiveness, broad search terms were applied. This resulted in the following search query: *sleep AND (consumer OR wearable OR cst OR "Consumer Sleep Technology") AND ("Machine learning" OR AI OR "Artificial intelligence" OR "Deep Learning").* As seen in Fig. 2 there was a spike of interest in 2017. Therefore only records from 2017 to 2021 will be reviewed. This search was performed September 24, 2021.
- **Identification:** A total of 248 records were identified. Based on a screening of title alone 96 records were identified as potentially relevant.

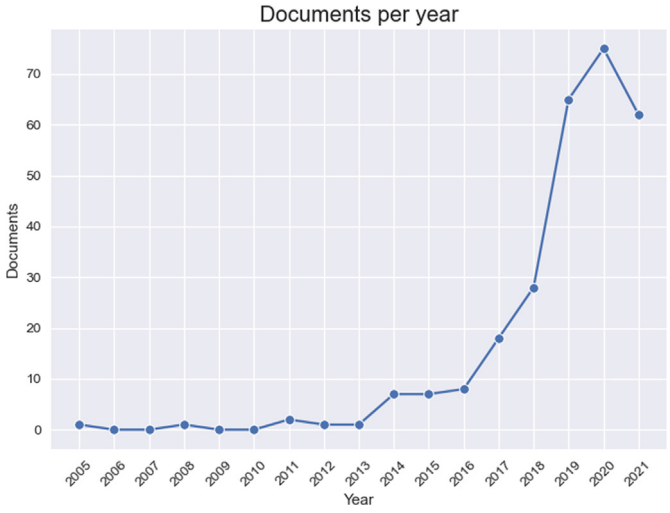

Fig. 2. Documents related to search query per year.

- **Screening:** This was done in two passes. First applying the inclusion criteria on the abstract and skimming of the text and then again after a full read. On the first pass 36 records were identified and on the second pass 27 records were chosen for synthesis. The records were selected by the first author and in case of doubt the other authors were consulted.
- **Inclusion criteria:** studies were included that met the following criteria: 1) sleep classification was performed using ML or DL; 2) devices used to capture physiological data were either wearables or had the potential to be wearable; 3) ground truth labels came from AASM rules; 4) English language and peer-reviewed journal, conference or workshop.
- **Data extraction:** 27 records were chosen for review. From each study the following information was extracted, if available: device and sensors, amount of data used, availability of datasets, type of ML/DL model, raw signal or feature engineering approach, modality of model, number of sleep stages classified, and the model performance.
- **Synthesis of results:** Due to the vast difference in the datasets used, lack of data availability and difference in methodologies a meta-analysis is not appropriate, hence a systematic qualitative review was conducted.

Initially records were excluded based on title alone. This left 96 records. Afterwards the abstract of each record was read and the text was skimmed. This left 36 articles to be reviewed in depth. After a full reading 27 articles were chosen for synthesis. These comprised of 16 conference articles and 11 journal articles [26]. is a review about wearable general for healthcare. The only sleep related section was about the Cole-Kripke algorithm so it was deemed not relevant to this study. This process can be seen in Fig. 3 along with exclusion criteria.

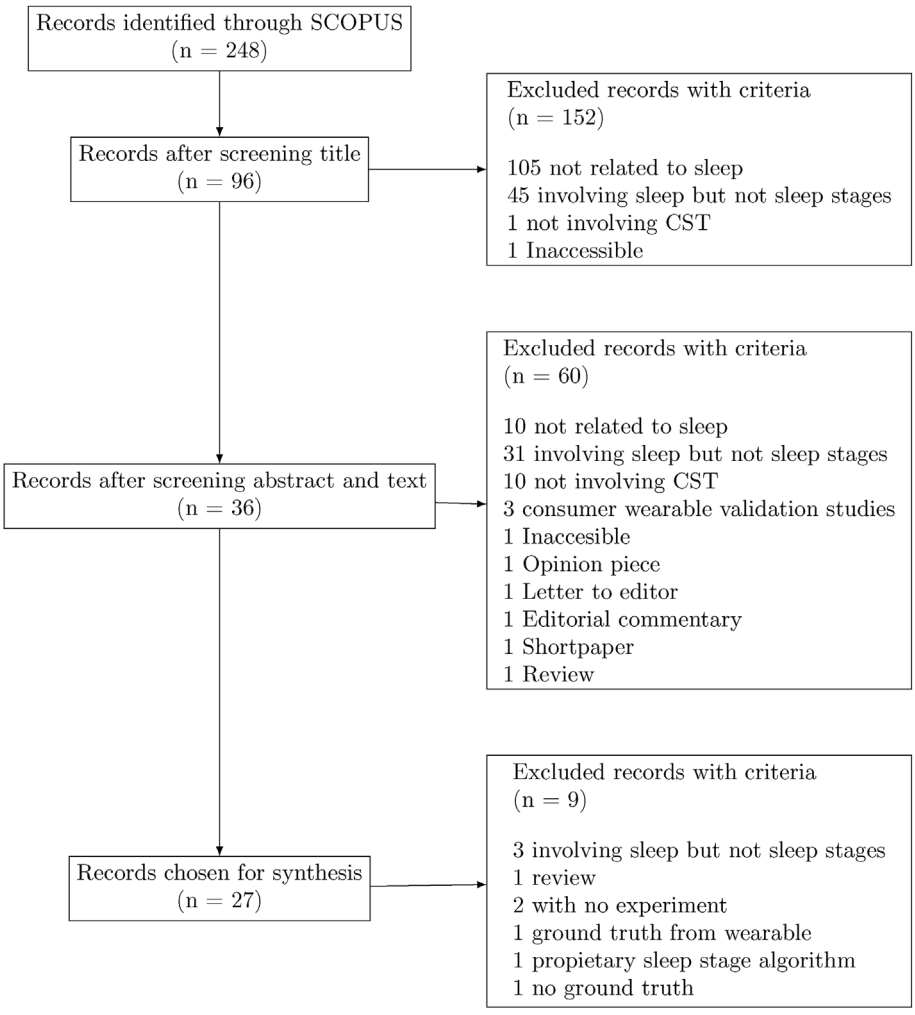

Fig. 3. Flowchart of exclusion of records and exclusion criteria. From an initial 248 records found from the search query, 27 were chosen for to review.

4. Devices & datasets

This section will summarize what CSTs have been used in the literature, the amount of data gathered and datasets available.

4.1. Devices

There are many different CSTs that exist both on the market and off the market. A comprehensive review of these can be found in Ref. [17]. The ones used in the 27 articles chosen for synthesis can be seen in Table 3. The devices' name, physical location, commercial availability, sensors in the device and their sampling frequency have been noted. The most common type of CSTs are the wrist-worn devices with 17 out of the 32 devices used in the studies. They also often contain multiple sensors which allow for multi-modal classification of sleep stages like the finger-worn OURA ring [27] which has a Photoplethysmogram (PPG), an accelerometer and skin temperature measurements. Not every study specified all the details about the chosen device. For an example [28] only specified that they used a finger-worn PPG but specified sampling frequency. On the other end [23] used a Samsung Gear S2 smart-watch which contains multiple sensors but only used PPG which is why only PPG was reported as a sensors.

Some articles mention how much data they gathered and some only mention the number of participants. 15 studies only used data they gathered themselves, 8 studies only used dataset from other studies and 4 did both. The additional datasets are described in Table 2 [30]. has an asterisk as they did not use a PSG to score sleep stages but instead used a single channel EEG. and scored according to rules from AASM.

4.2. Datasets

One of the crucial aspects that has allowed ML to be as influential as it is in a field like image recognition is the amount of data that has become available, especially labelled data. This process is very resource exhaustive in the sleep classification domain as manually labelling requires expert coders and PSG recordings. A summary of public datasets used in the 27 studies can be seen in Table 2. The largest studies being Multi-Ethnic Study of Atherosclerosis (MESA) and The Sleep Heart Health Study (SHHS) are both restricted access, which means you have to apply to use them while datasets on PhysioNet and Zenodo are open access. Locating the actual datasets was a challenge since e.g. Ref. [50] mention that they used Technische Universität Darmstadt (ICHI14) but referred to it as TUD. Furthermore the only reference to the dataset was [53] which is the article that introduced the dataset. Unfortunately [53] linked to a website that did not exist anymore as it had changed since 2014. This issue can be circumvented by attaching a more permanent identifier to a dataset like a DOI which many have done.

A hyperlink to the dataset location has been included for all datasets except Whitehall II and Fitabase-fitbit PPG which just link to their respective websites. There were 3 datasets which could not be located; Fitabase-PPG used in Ref. [43], the CONTEXT study used in Ref. [50] and a dataset from University of Pensylvania used in Ref. [51]. The authors were also unable to learn more about the Whitehall II study beyond that they had participants self-reporting their sleep/wake status. All studies except Whitehall II mentioned in Table 2 are PSG validated.

Even though there exist large studies like MESA they often lack diversity of sensor modalities for CSTs. The MESA study includes a wrist-worn actiography but if another modality like a PPG is desired the pool of data is lacking. Since there also exist a plethora of different CSTs as seen in Ref. [17] researches often end up collecting their own datasets. Unfortunately most of these are not made publicly available. Table 1 show an overview of the amount of data the chosen studies have used. It is quickly evident that the studies fall into three categories; 15 articles that only use data they gathered themselves, 8 that only used publicly available data and 4 who did both. It is also evident that only [42] made the data they gathered available. An issue that makes it challenging to compare datasets across studies is how they report their data. Since most are not available it matters how they describe the data. Some studies only report the number of subjects/participants. Others measure it in number of nights/recordings/days. Others use hours or epochs to describe the data. Also as seen in Table 1 not every study uses PSG validated data [38]. as an example made subjects self report when they went to bed and when they were awake and used that as ground truth for awake/sleep classification. The way additional datasets were used also differed in the studies [44, 42, 43]. trained a model on their own data and tested against an additional dataset [45]. did it the other way around by training on the additional dataset and testing against a part of the additional dataset and their own dataset.

5. Classification approaches

This section will describe what types of algorithms were used in the 27 selected articles and how they were used. A summary can be seen in Fig. 4. The full table of every article can be found in the appendix Table A1. The authors were unable to determine which model in Ref. [29] performed the best, therefore their models have been left out of the table to not crowd it as they tried Conditional inference tree, Random Forrest (RF), Logistic model trees, Naïve Bayes, Nearest shrunken centroids and SVM. The rest of their approach was included in the table. Table 4 shows the taxonomy of the sleep stages and how they relate to the AASM stages. The presented attributes are; How many sleep stages were used; if the algorithm was based on ML or DL; if data was from one sensor or multiple modalities; what algorithm was used and what sensors were used. Some papers [47, 51, 27, 39, 44, 34, 32, 42] tried multiple sleep stage combinations and therefore the best performing algorithm was chosen for each case. This causes the summary to add up to more than 27 and should instead be read as X out of 27 papers used this approach in their study. For the specific algorithm they have been simplified for the summary, e.g. Ref. [45] used a Complex-valued unsupervised CNN but this is just counted as a CNN.

2-stage sleep classification is the most common with 14 of the papers they have used that. There is an even split between 4 and 5 stage classification with 3 stage being the least common. Unimodal approaches are by far the most common with 18 papers using this approach. The most common sensor are accelerometers which make sense as actiography is a well researched area. Another popular approach is sensors measuring heart activity through PPG (7), ECG (4), Ballistocardiography (BCG) (3). There are also 5 studies using EEGs and while an EEG is typically not very mobile the studies make use of 1-channel EEG like [48] or a mobile EEG like [32,46]

There seems so be an even split between using ML (14) and DL

Table 1
Summary of sleep data used in the 27 articles chosen for review.

Author	No. of participants	Data gathered	PSG validated	Dataset available	Additional datasets	Total data
[28]	10	10 nights	Yes	No		
[29]	23		No	No		
[30]	23		No*	No		
[31]	22		Yes	No		
[32]	17	165 Hours	Yes	No		
[33]	25	24029 epochs	Yes	No		
[34]	19		Yes	No		
[35]	25		Yes	No		
[36]	10	10 nights	Yes	No		
[37]	5		Yes	No		
[38]	3	21 nights	No	No		
[39]	50	50 nights	Yes	No		
[27]	106	440 nights	Yes	No		
[40]	10	123134 epochs	No	No		
[41]	26		Yes	No		
[42]	31		Yes	Yes	MESA	219 subjects
[43]	25	51 nights	Yes	No	MIT-BIH, Fitabase-Fitbit PPG	143 nights
[44]	17	2 weeks	No	No	ICHI14	59 subjects
[45]	32		Yes	No	SHHS	449 subjects
[46]					MIT-BIH	39 nights
[47]					MESA	1743 nights
[48]					Sleep-edfx	197 nights
[49]					Sleep-edfx	36 nights (20 subjects)
[23]					CAP, MIT-BIH, UCD	410 Hours
[50]					CONTEXT study, ICHI14, Newcastle PSG	145 nights
[51]					Newcastle PSG, University of Pensylvania, Whitehall II	156 subjects
[52]					UCD, MIT-BIH	

Table 2
Summary of datasets from other studies used in the 27 articles chosen for review. Fitabase is the only dataset which could not be located hence the missing accessibility. Whitehall II is a massive study spanning decades and the amount of participants in their sleep study was unassessable.

Name	Author	Accecability	DOI	Platform	Additional sensors	Subjects	Records
Multi-Ethnic Study of Atherosclerosis (MESA)	[54,55]	Restricted		National Sleep Research Resource (NSRR)	Accelerometer	2237	
MIT-BIH Arrhythmia Database	[56]	Open	10.13026/C2F305	PhysioNet		47	48
Fitabase-fitbit PPG	[57]			Fitabase	PPG	24	24
Technische Universität Darmstadt (ICHI 14)	[53]	Open		Technische Universität Darmstadt	Accelerometer	42	45
The Sleep Heart Health Study-1 (SHHS-1)	[58]	Restricted	10.25822/ghy8-ks59	National Sleep Research Resource (NSRR)		6441	
The Sleep Heart Health Study-2 (SHHS-2)	[58]	Restriced	10.25822/ghy8-ks59	National Sleep Research Resource (NSRR)		3295	
Sleep-EDF Database Expanded	[59]	Open	10.13026/C2X676	PhysioNet			197
CAP Sleep Database	[60]	Open	10.13026/C2VC79	PhysioNet			108
St. Vincent's University Hospital/University College Dublin Sleep Apnea Database (UCD		Open	10.13026/C26C7D	PhysioNet		25	25
Newcastle PSG	[61]	Open	10.5281/zenodo.1160410	Zenodo	Accelerometer	28	28
Whitehall II		Restricted		University College London (UCL)	Accelerometer		

Table 3
Table of the CSTs used in the literature, where they are placed during sleep, what kind of sensors were used in the studies, and the sampling rate of the sensors in Hz if they described it.

Author	CST	Device location	Commercially available	Sensor	Hz
[43]	Dozee	Under matress	Yes	BCG	250
[28]	Not specified	Finger		PPG	128
[46]	Headband EEG	Head	No	EEG	
[47]	Actiwatch Spectrum	Wrist	Yes	Accelorometer	
				ECG	
[45]	SensEcho	Vest	No	ECG	200
				accelerometer	25
[49]	Muse	Head	Yes	EEG	220 (100)
[23]	Samsung Gear S2 smartwatch	Wrist	Yes	PPG	25
[50]	AX3	Wrist	Yes	Accelorometer	
	HedgeHog	Wrist	No		
	GENEActiv	Wrist	Yes		
[29]	Fitbit Charge 2	Wrist	Yes		
[31]	The Philips Actiwatch 2	Wrist	Yes	Accelorometer	32
[30]	Fitbit Charge 2	Wrist	Yes		
[32]	Ear-EEG	Ear	No	EEG	1200
	Scalp-EEG	Scalp			
[33]	Eye Mask	Face	No	ECG	
				EOG	
[34]	Zephyr BioHarness 3	Chest	Yes	ECG	1
	MSR 145B3	Wrist & Ankle	Yes	Accelerometer	51.2
[35]	WhizPad	Under matress	Yes	BCG	
[36]	Wave	Near bed	No	Passive infrared sensor	
	Galaxy Note 8	Near bed	Yes	Microwave sensor	
[37]	Their own	Under matress	No	ballistocardiography	100
[38]	Empatica E4	Wrist	Yes	EDA	1
				PPG	
				Accelorometer	
[39]	Samsung band	Wrist	Yes	Accelerometer	20
				PPG	20
[27]	Oura	Finger	Yes	PPG	125
				Accelerometer	50
				Temperature	
[51]	GENEActiv	Wrist	Yes	Accelorometer	
[44]	wActiSleep-BT	Wrist	Yes	Accelerometer	60
	EZ430-Chronos	Wrist	Yes	Accelerometer	60
[40]	GT3X (Actilife, USA)	Wrist	Yes	Accelerometer	100
[41]	Zephyr BioHarness 3	Chest	Yes	ECG	1
	MSR 145B3	Wrist & Ankle	Yes	Accelerometer	51.2
[42]	Apple Watch	Wrist	Yes	Accelerometer	50
[48]				Single channel EEG	200
[52]				EEG	

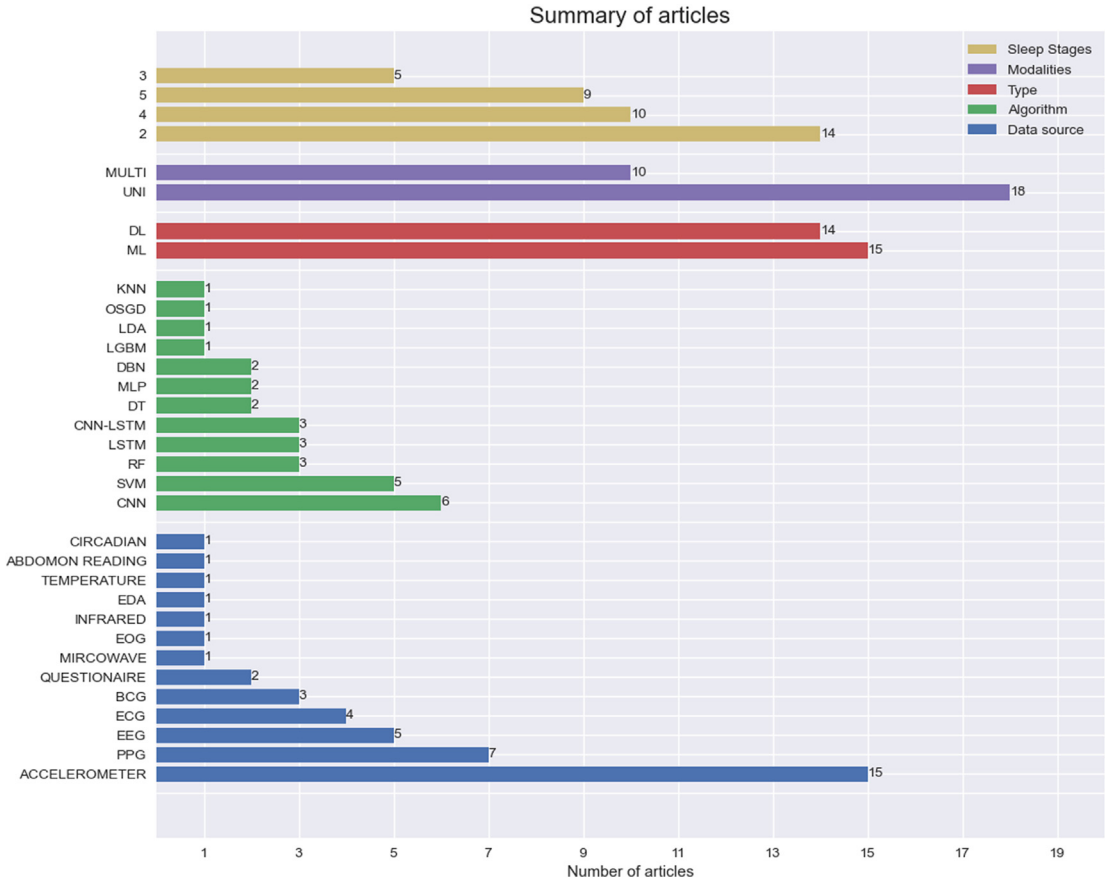

Fig. 4. Summary of approaches performed amongst the 27 articles. Since some studies perform multiple sleep stage classifications the total of each group does not sum up to 27. There are 5 groups on the figure describing the number of sleep stages they tried to classify, the modality of the algorithm, if it was a ML or DL model, what class of algorithm was used and where the data used in the models was from.

Table 4
A summary of the different terminology for talking about the different sleep stages. It begins from 5 sleep stages and shows which classes are collapsed as the number of sleep stages descend.

Sleep stages	Terminology				
5	W	N1	N2	N3	R
4	Wake		Light	Deep	R
3	Wake		NREM		R
2	Wake		Sleep		

(15). For ML the most common models are SVM (6) and RF (3)/ Decision Trees (DT) (2) which are related models. For DL, CNN (6), LSTM (3), and architecture which combines them CNN-LSTM (3) are popular choices.

5.1. Modalities

As seen in Fig. 4 unimodal approaches where only one sensor is used is the most researched approach. It can make it easier to understand how well that specific sensor performs sleep classification. On the other hand an increase in modality can make the performance better. There are two ways which modalities have been combined amongst the 27 papers. The first is by combining data from multiple sensors. This has been done in Ref. [38] where they use the Empatica E4 wrist device that has various sensors amongst which they use PPG, Electrodermal activity (EDA) and the accelerometer. They extract features for the sensors for a given 30 s segment of data like the mean of EDA, mean heart rate, standard deviation for the Blood Volume Pulse (BVP) and mean of the accelerometer signal. They feed these into an SVM to classify sleep/ wake. Their best performing model was the one that combined all four features. Similarly [47] extract HRV and accelerometers features and show how combining them increases performance as they go from 2 stage sleep classification to 3, 4 and 5 stage sleep classification. At 2 sleep stages, the accuracy for the best performing models for multi (84.4%), actiography (84.9%) and HRV (79.5%) does not differ much. At 5 sleep stages, multi (63.7%) and actiography (55.6%) and HRV (56.9%) differ significantly indicating that there is additional information in the different signals that provide a richer signal when combining them.

5.2. Classical machine learning vs Deep Learning

As mentioned before the classical approach is using hand crafted features as input for the model [23]. is an example of this approach. They calculated 43 features from HRV for their models. These features are from the time domain like mean and standard deviation, the frequency domain with power of the various frequency bands, and some non linear features. This resulted in a feature vector of length 43 which was used as input in an SVM to

predict sleep/wake. Similarly [51] had a similar approach where they calculated 12 features for each axis of the accelerometers which resulted in a 36 feature vector that was fed to a RF. Another way of using features can be seen in Ref. [29] where they in tandem with an accelerometers and PPG also use demographical features like the sex, age, and subjective sleep quality from a questionnaire of the participants. DL on the other hand allows the opportunity to use move past feature engineering and use the raw signal instead where a model will learn internal features to represent the signal. This was done in Ref. [40] where they tested both approaches. For the classical approach they extracted 36 features from an accelerometer across the time domain and frequency domain and used as input in various models including RF and a feature based CNN to classify sleep/wake. The other approach was to use the x, y and z axis signal from the accelerometer as input for their CNN. This increased the performance from and accuracy of 84.14% with the feature based approach to 89.65%. Out of the 14 papers that used a DL model, 9 used some form of the raw signal as input [40, 37, 50, 52, 43, 49, 41, 34, 33]. In most cases the raw signal was prepossessed initially by either some band filtering or normalizing before usage [50]. made a CNN that took both an acceleromoter axis and its corresponding Fast Fourier Transform coefficients as features as input, thus combining raw signal and handcrafted features [33]. transformed EEG and EOG signal into a spectogram image and used a CNN called MobileNetV2 which was pretrained for image classification tasks [34,41]. combined ML and DL as they features generated by a Deep belief network (DBM) as input in a RF model. A DBM is essentially a model that tries to probabilistically reconstruct the given input which in this case was 30 s accelerometers recording. The hidden layers of the DBM can then be used as features in classification model.

5.3. Sleep stage classification performance

Evaluating model performance and comparing models approaches across papers is not a straight forward task. First of all there does not seem to be agreement of how this should be reported in the field. We observed that the most common metrics were accuracy, F1, sensitivity, precision, Cohens Kappa and a confusion matrix of the sleep classes. By far the most common is accuracy, but this metric has a major drawback as sleep stages are inherently uneven. In wake/sleep classification the distribution favours sleep heavily as participants usually are asleep most of the night. Therefor by just classifying everything as sleep the model will achieve a high accuracy. For a full REM cycle of adult sleep the distribution of time in each sleep stage approximates 5% N1, 50% N2, 20% N3, and 25% REM [62]. This results in N1 occurring the least and therefore the same problem arises. To account for this inherent imbalance in the distribution of training data there exist multiple metrics to describe model performance. Among the common ones seen in the 27 studies were sensitivity, precision, F1- score, Cohens Kappa and confusion matrices. We will provide a brief and intuitive understanding of these and what they show. Sensitivity is the fraction of true positives that were correctly classified and precision is the fraction true positives of everything classified as positive. F1 is a the harmonic mean of precision and sensitivity. These three range from 0 to 1 with 1 being highest score. Cohens Kappa, often written as κ, is a measurement that also takes into account that correct classification could occur by chance and ranges from -1 with complete disagreement, to 1 being complete agreement. The confusion matrix is simply a matrix that shows all the models classification compared to what the ground truth is and is reported in either percentages or total number classified. The confusion matrix has the advantage that is shows which classes the model struggles with and how they are misclassified. Another aspect is how the authors choose to validate their models. It is common practice in the ML field to reserve a portion of a dataset which has not been used to training the algorithm as a test set. There is no rule as to how much training data to use but a typical split is 80% for training and 20% for testing. This approach has the drawback that the model is only run once and the variability of model performance is not captured. To combat this problem another validation methodology called K-fold Cross Validation (CV), where the dataset it split into K partitions of equal size and the model is trained from scratch on the rest of the data. This is repeated for each K. So a 5 fold CV means that the dataset is split into 5 equally large parts, and the model is training on 80% of the data and tested on 20% of the data 5 times. This results in an average performance of the model with the variance shown even though the variance is not always reported. A different variation of CV is Leave-One-Out CV where in the case of sleep stages either a subject or recording is left out training and used for testing. If a subject is left out it shows how well a model performs on a subject it has not been before while the other methods show how well it performs when it has seen some sleep data from the subject. As it has been referenced in different ways in the studies e.g. LOOCV, LOOV, LOSO, we choose to abbreviate it Leave-One-Subject-Out (LOSO) to be explicit. Lastly some also use a completely different data set to measure how well the model generalises to completely unseen data, which often causes a drop in performance compared to the test set. All of these different choices of measuring a models performance can make it difficult to compare models and conclude which specific model is the best.

For sleep/wake classification it can be seen that out of the 5 studies that have above 90% above accuracy 3 of them combine PPG with an accelerometer. The model with the highest accuracy, 96%, is by Ref. [27] but given their F1-score of 0.78 which indicates that it overestimates one of the classes. Depending on what the objective is a higher F1-score is preferable on the cost of a loss in accuracy [40]. which only uses accelerometer sensor but uses the raw signal with a CNN-LSTM also performs reasonably with an accuracy of 88.77% and sensistivy and precision above .9. The only EEG model for 2 stage sleep classification is also the best performing unimodal model with an accuracy of 95.2 and a κ of 0.83. It seems that the unimodal approach of accelerometer or heart rate are performing similarly.

For 3 stage sleep classification is seems that [37] vastly outperform the other approaches with their BCG based CNN-LSTM model which uses the raw signal. It was initially trained on sleep posture labels of subject and then fine tuned to sleep stage labels. They do not report a total accuracy but report their confusion matrix which results in accuracies for wake (95.3%), NREM (93,1%) and REM (84%) [34, 41]. seem to be the other approach where they take features from an accelerometer and use a RF model and achieve N1 of 37.3% and 32.9% respectively. The only heart rate based approach [47], does not seem be able classify N1 nor N3.

For 4 stage sleep classification [36] has by far the best performance reaching an accuracy of 99.65% by combining contact free microwave sensor and an infrared sensor with a K-nearest neighbour (KNN) model. These results are quite incredible but are difficult to reproduce as they have not released neither their training data nor their code. They are also unclear about their definitions of the different sleep stages as they in one instance write " … L: Light sleep (non-REM sleep stage 1); D: Deep sleep (non-REM sleep stage 2 and 3)." and in another" … light sleep (non-REM 1 and 2), and slow-wave sleep (non-REM 3)." The seconds best model was [34] RF model with an accuracy of 90%. They use handcrafted 36 features from two accelerometers and compared this approach to DL features generated by a DBM. Their handcrafted features performed best but the DL features achieved an accuracy of 83.2%. The two EEG based models [33, 46] performed similarly with accuracies of

86.72% and 85.5%. They both used DL models but [33] used the spectrogram images of EEG and EOG signal. Their spectrogram included 15 s before and after a given 30 s epoch. The rest use a combination of heart rate sensors and accelerometers and achieve an accuracy between 70.3% and 80.75% [47]. is the only that seemed to be unable to do 4 stage sleep classification as it has an accuracy of 4% in the deep sleep stage.

For 5 stage sleep classification, accuracy is not a good metric. It is evident that the challenge lies in differentiating N1 and N2. The difference in accuracy between the top 2 [52], 87% and [32], 85.9% Is only 2.9% but their difference in N1 is 60.05%. Likewise the second highest N1 accuracy is [44] with 69.45% but with a significant drop in all other sleep stages. The two most successful approaches that achieve the highest N1% while not sacrificing the other sleep stages are [49,52] with N1 of 80.2% and 41.54% respectively. They have in common that they both use a CNN based arcitecture that takes the raw signal instead of feature engineering from an EEG.

†

6. Discussion

6.1. Potential areas of interest for improving classifiers

There seems to be untapped potential in the area of using raw signals with DL models for sleep classification. There are currently only six studies that have delved into this approach. The initial findings are that they perform very well and are amongst the best performers in 3 and 5 stage sleep classification. This also makes sense intuitively, as with more sleep stages the complexity also increases, especially differentiating between N1 and N2. Transforming the signal into a spectogram as [33] did and turning the task into a image classification task is an interesting approach with a lot of potential as this is a field where DL has been performing exceptionally well. This also allows to use neural networks that are pretrained on other visual tasks and fine tune them for sleep. A variation of this could be seen in Ref. [37], which used a raw BCG signal but was pretrained on a different task and fine tuned for sleep. There are more complex architectures designed specifically for raw signals that could be explored further like WaveNet [64].

To properly be able to utilize the power of DL models, large amounts of data are needed. As seen in Table 1 only one study made their data available. We strongly urge researchers in this field to follow the FAIR principles [65] and make both the labelled sleep recordings from PSG, their chosen CST, and the models they used available on easily accessible repositories such as PhysioNet or Zenodo, which already hosts some labelled sleep recordings as seen in Table 2. It would also make it possible to recreate or validate the studies as that is currently impossible without access to their data or software.

Currently most publicly available datasets do not contain additional wearable sensors. It would be valuable to be able to harness these datasets for models to be used with wearables. Especially sensors that measure heart rate are an obvious choice as a PSG includes ECG and many wearable have PPGs. Recent studies also point to HRV being the best sole indicator of sleep stages [66] which has shown great results with DL [67]. An example of this could be seen in Ref. [43] where a model was pretrained on ECG and PPG data but used with a BCG sensor. Similarly [68] also trained a DL model on HRV features from an ECG and finetuned the model with HRV features from a PPG. A variation of this approach could be to transform the PPG signal into an ECG, signal which has been done in a pilot study by Ref. [69]. Another unexplored avenue to increase the amount of data used for training, is to use synthetically generated data. This has been seen in the field of eye movement classification which has certain parallels to sleep stage classification. Both fields try to classify events based on a biometric signal according to a ruleset, and manually labelling data is very time consuming and requires training. To address the problem of data availability an Recurrent Neural Network (RNN) based model was trained to generate synthetic signals of eye movements and then a separate neural network was trained on the synthetic data to classify the movements [70].

Another way to improve the models is by combining modalities to create a richer signal. This was seen in Table 5b and Table 5d where performance increased, especially with number of sleep stages. Most commonly seen in the reviewed papers was to combine accelerometer with a heart based sensor, but it was also shown that combining demographical information into the model increased performance. There are many stages that modalities can be combined at; signal, feature, score, rank, or decision level [71]. All the attempts described earlier are combined at the feature level where features from different modalities are extracted and merged into a single feature vector and fed to a model. We believe these different stages of combining modalities are worth exploring to investigate how much the models performance can be improved. Especially a more advanced method such as a fusion network [72] are interesting. An example of this in the context of sleep stage classification, is a neural network which takes a 30 s segment of signal from a PPG as input in one channel and the corresponding segment from an accelerometer as input in a second channel with the last layers of the network concatenating their output before classification. These are not restricted to only two modalities or to using a CNN architecture. A simplified depiction of this architecture is shown in Fig. 5. In addition to increasing modalities there is also the aspect of individualising the models. This can be done by investigating the effect using a model trained on a general population and finetuning it for an individual participant or demographic [73].

6.2. Designing for interactive sleep intervention

There exists a large of body of work detailing how sleep can be improved which is beyond the scope of this study. Nonetheless there seems to be a focus on sleep hygiene [4] where preventive measures are taken by changing lifestyle and habits like regular exercise, avoidance of caffeine, nicotine, and alcohol. As CSTs improve and Internet of Things becomes more advanced and widespread it will become possible to measure the sleep cycle at home and intervene during sleep instead of only using preventive measures. An example of this is wearing a smartwatch or having a BCG sensor under the bed while sleeping, that classifies sleep stages in real time and adjusts the physical properties of the environment like lighting, temperature, and ambient sounds of a room. In the former we have emphasized the technical potential for constructing ML infused systems to facilitate such changes in real time to support users' sleep. However, developing successful systems is also a matter of designing the user interfaces in a way that enables and engages interactions. Given the recent advances and increased availability of ML infused systems it is necessary to consider how to design these. This is because ML enables a higher level of system autonomy compared to systems in which the users traditionally have had full control. Van Berkel and colleagues differentiate between *intermittent*, *continuous* and *proactive* ML systems, with the most recent advancements pointing towards users interacting with ML systems that act like real-time collaborative partners, i.e. continuous ML interaction rather than conventional intermittent interaction, described as turn-taking [74]. Looking a bit further ahead we should also expect more proactive systems to appear, that is, ML systems that act more autonomously and not necessarily based on explicit user input, e.g. ML systems

Table 5
Summary table of best performing sleep classification models from each study. Describes the inputs of the model, algorithm and validation methods used, and performance metrics. Accu. describes the accuracy. Validation describes the respectively the training/test split, type of CV denoted by number of splits or LOSO. The independent dataset is a dataset from a different source than training data. MCCV is a Monte Carlo CV where a dataset is split randomly N times into training and testing. †denotes the usage of the raw signal instead of engineered features. (B) is a balanced accuracy which is different from regular accuracy. (W) is a weighted, F1, sensitivity and precision. Acc is an accelerometer. Temp is skin temperature. Circ. is a mathematical modelling of the circadian rhythm.

(a)

2-stage sleep classification: wake, sleep

Author	Signal	Algorithm	Validation	Accu. (%)	F1	Sensitivity	Precision	Specificity	Kappa	Wake (%)	Sleep (%)
[27]	PPG, acc, temp, circ.	LGBM	5 CV	**96**	0.78	0.8074		**0.9815**			
[32]	EEG	SVM	10 CV	95.2					**0.83**		
[38]	PPG, acc, EDA	SVM	80/20		**0.93**						
[63]	BCG	DT	10 CV	90.9		**0.957**		0.775		**77.5**	**95.72**
[42]	PPG, acc	SVM	50 MCCV 70/30	90.1					0.449	59.6	93
[40]	acc	CNN-LSTM †	10 CV	88.77		0.9296	0.9039				
[47]	acc	CNN	80/20	84.9	0.883	0.938	0.847	0.671	0.63	67	93
[28]	PPG	SVM	70/30	81.1	0.8174	0.8106	**0.9937**	0.8250			
[23]	ECG	SVM	80/20	80.1 (B)		0.786		0.816	0.801		
[30]	PPG, acc, questionnaire	DT	LOSO			0.848	0.614				
[51]	acc	RF	75/15	78.76	0.7393					58.93	89.66
[50]	acc	CNN †	Independent data set	77		0.83		0.5			
[31]	acc	LSTM †	80/20	67.7		0.377	0.907				

(b)

3-stage sleep classification: wake, NREM, REM

Author	Signal	Algorithm	Validation	Accu. (%)	F1	Sensitivity	Precision	Specificity	Kappa	Wake (%)	NREM (%)	REM (%)
[37]	BCG	CNN-LSTM †	LOSO							**95.3**	**93.1**	**84**
[32]	EEG	SVM	10 CV	**90**					**0.8**			
[39]	PPG, acc	LDA	LOSO	84	**0.85 (W)**	**0.84 (W)**	**0.85 (W)**		0.67			
[47]	ECG, acc	LSTM	80/20	76.2	0.679	0.688	0.722	**0.856**	0.584	75	84	48
[42]	PPG, acc	MLP	20 MCCV 70/30	72.3					0.277	60	65.1	65

(c)

4-stage sleep classification: wake, light sleep, deep sleep, REM

Author	Signal	Algorithm	Validation	Accu. (%)	F1	Sensitivity	Precision	Specificity	Kappa	Wake (%)	Light (%)	Deep (%)	REM (%)
[36]	microwave, infrared	KNN	LOSO	**98.65**						**96.5**	**98.3**	**99.5**	**99.6**
[34]	acc	RF	10 CV	90						84.4	92.3	90.2	87
[33]	EEG, EOG	CNN †	68/32	86.72						85.2	87.17	82.87	89.30
[46]	EEG	MLP	LOSO	85.5						94.8	81.9	88.9	79.2
[45]	ECG, acc, adb. reading	LSTM	Own dataset	80.75					**0.69**	87.2	75.5	90.8	88.8
[27]	PPG, acc, temp	LGBM	5 CV	79	**0.78**								
[39]	PPG, acc	LDA	LOSO	77	0.76 (W)	**0.76 (W)**	**0.76 (W)**		0.58				
[43]	BCG	CNN-LSTM †	80/20	74	0.73	0.74	0.73			70	75	76	64
[47]	ECG, acc	LSTM	80/20	70.3	51.9	54.0	57.9	87.4	53.8	77	80	4	55

(d)

5-stage sleep classification: wake, N1, N2, N3, REM

Author	Signal	Algorithm	Validation	Accu. (%)	F1	Sensitivity	Precision	Specificity	Kappa	Wake (%)	N1 (%)	N2 (%)	N3 (%)	REM (%)
[52]	EEG	CNN †	LOSO	**87**					**0.8**	88	**80.2**	**94.05**	90.07	85.02
[32]	EEG	SVM	10 CV	85.9					0.79	83.77	20.25	89.48	86.12	81.78
[34]	acc	RF	10 CV	84.5						86.8	37.3	90.5	90.2	89.6
[48]	EEG	SVM	5 CV		**0.85**		0.836	0.847						
[49]	EEG	CNN †	95/5	81.72	**0.76**	0.77	**0.76**			**96.16**	41.54	82.87	81.31	82.61
[41]	acc	RF	10 CV	80.7						90.3	32.9	83.1	**91.3**	**91**
[44]	acc	OSGD	Unclear	73.2						74.78	69.45	70.29	58.22	58.22
[47]	ECG, acc	CNN	80/20	63.7	0.399	0.43	0.471	**0.887**	0.563	80	2	79	6	49
[51]	acc	RF	75/15							59.82	02.49	71.28	18.13	7.95

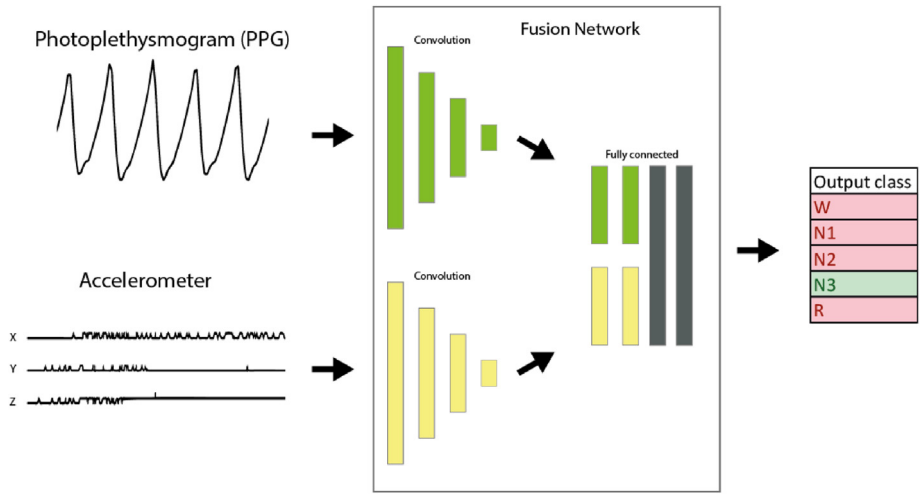

Fig. 5. Depiction of a simplified fusion network. There are two input channels for the neural network that take the raw PPG and accelerometer signal. A convolution structure is used to learn a representation of the signals which is combined in the fully connected stage and used to predict sleep stages.

that act on sensor data rather than input consciously provided by a user [74]. Related to the notion of proactive ML systems, Janlert and Stolterman highlight the concept of *agency* in interactive systems [75]. Janlert and Stolterman note that user perceived interactivity is a matter of balancing between too low system agency (system being perceived as too predictable or "pliant") and too high system agency (system being perceived as too unpredictable or "having its own will"). Either of these extremes represents a risk of the system not enabling and engaging user interaction [75].

The unique challenges in designing the user experience (UX) of ML infused systems has been discussed for years in the Human-Computer Interaction research community, cf [76]. The overarching question being: *To what extent can UX designers rely on well-established design principles to shape user interfaces in a way that* supports *the users' mental models of ML systems' behaviour?* Amershi and colleagues [77] provides a synthesis of previous efforts in guideline development and suggest a range of 18 design principles for UX designers to follow when designing ML infused interactive systems. While this work is commendable, their guidelines have been critiqued for essentially not dealing with challenges unique to the design of interactive ML infused systems [76]. What [76] propose instead is to consider ML systems at four levels with the first level not representing unique design challenges. Level 1 is defined as an ML system with a fixed capability (e.g. face detection) and limited system output (e.g. face recognized or not) while level 4 is defined by evolving system capabilities (e.g. search engines that learn from new data after deployment) and the possibility of infinite system outputs (e.g. result suggestions in a search engine) [76]. Our case of interactive ML based noise masking may be classified between level 1 and level 2 given the fixed nature of such a systems' capability (to classify sleep stages and invoke noise masking) and the output complexity of the ML algorithm being constrained to a finite number of sleep stages. According to Ref. [76] systems like that envisioned here may not necessarily invoke the need to develop and utilize design guidelines uniquely applying to interactive ML infused systems. Yet, although the output of an ML algorithm may be constrained to a finite number of sleep stages, it will lead to intervening instructions on, e.g. changing the ambient sounds (noise masking), lighting, temperature etc. Such interventions represent a higher level of output complexity.

Additionally, dealing with a system that interacts with the user being in an unconscious state (sleeping) and based on real-time sensor data represents an avenue of work that needs to be further explored. This is because there is no visual, auditory or tactile interface through which the user provides input and receives feedback, i.e. the system is proactive [74]. We argue that the type of interaction at play during sleep must be *implicit* rather than *explicit*, meaning that the user can interact with the ML system in way that does not require the users' attention. According to Janlert and Stolterman such unattentiveness would bear more resemblance to automation rather than "genuine interaction" [75], but it can also be argued that the type of system discussed here represents a case of "genuine implicit interaction", simply due to the fact that you are sleeping, i.e. unconscious while the system not only registers your physiological input, but it also reacts towards this. This is a different case than implicitly interacting with e.g. an automatic sliding door in a building on your way to a meeting. You are aware of the door and implicitly assume that it will open, but your main intent is to get to the meeting and are therefore less attentive towards the door. Such genuine implicit interaction leads to a wealth of questions worth exploring, e.g. how do users experience a system that they implicitly interact with during a state of unconsciousness? How and when should users explicitly interact with the system in order to adjust the classification model to be more precise? To what extent are users aware of the system making the correct classifications of sleep stages, i.e. do they even notice if the ML algorithm makes mistakes? How can we strike a balance in perceived system agency such that it will lead to sustained usage of the system?

7. Conclusion

This work shows a systematic review of the current landscape of AI in sleep stage classification using CSTs. There is a trend of model performance increasing by combining modalities such as accelerometer with heart based sensors like a PPG, especially when moving from sleep/wake classification to 3 and 4 stage classification. For 5 stage sleep classification multi-modal approaches are less common. The majority of approaches reviewed used models trained on engineered features but there is a subset of studies that use the raw signal as input for DL models that perform

comparatively. We believe that the DL approach shows great promise and we discuss a future road map for how these DL models can be improved. A crucial aspect for this to be possible is for researches to make their PSG and CST sensor data and sleep labels available to fellow researchers due to the requirement of large data sets to apply DL. As these models reach maturity, new methods of sleep intervention, based on real time sleep classification, are possible. These interventions also open a set of challenges from an interaction design perspective, which are vital to consider when designing a system that operates while the user is not consciously aware of the changes that are happening.

Author contributions

The author S.D found the literature, synthesised the review, and wrote most of the paper. A.B helped frame the paper, define the scope, and wrote about the interaction challenges. T.D.N provided guidance in the machine learning aspects of the paper. All authors contributed and reviewed the final version of the paper.

Declaration of competing interest

The authors declare no competing financial or non-financial interests.

Acknowledgements

This publication was funded by SoundFocus ApS and Innovation Fund Denmark under grant number 1044-00079B.

A Appendix

References

[1] Medic Goran, Wille Micheline, Hemels Michiel. Short- and long-term health consequences of sleep disruption. Nat Sci Sleep may 2017;9:151–61.
[2] Krause Adam J, Simon Eti Ben, Mander Bryce A, Greer Stephanie M, Saletin Jared M, Goldstein-Piekarski Andrea N, Walker Matthew P. The sleep-deprived human brain. Nat Rev Neurosci jul 2017;18(7):404–18.
[3] Uehli Katrin, Mehta Amar J, Miedinger David, Hug Kerstin, Schindler Christian, Holsboer-Trachsler Edith, Leuppi Jörg D, Künzli Nino. Sleep problems and work injuries: a systematic review and meta-analysis. Sleep Med Rev 2014;18(1):61–73.
[4] Irish Leah A, Kline Christopher E, Gunn Heather E, Buysse Daniel J, Hall Martica H. The role of sleep hygiene in promoting public health: a review of empirical evidence. aug 2015.
[5] Basner Mathias, McGuire Sarah. WHO environmental noise guidelines for the European region: a systematic review on environmental noise and effects on sleep. Int J Environ Res Publ Health mar 2018;15(3):519.
[6] Stanchina Michael L, Abu-Hijleh Muhanned, Chaudhry Bilal K, Carlisle Carol C, Millman Richard P. The influence of white noise on sleep in subjects exposed to ICU noise. Sleep Med sep 2005;6(5):423–8.
[7] Attarha Mouna, James Bigelow, Michael M. Merzenich. Unintended consequences of white noise therapy for tinnitus - otolaryngology's cobra effect: a review. oct 2018.
[8] Ibáñez Vanessa, Silva Josep, Cauli Omar. A survey on sleep assessment methods. PeerJ may 2018;6(5):e4849.
[9] Patel Pious, Kim Ji Young, Brooks Lee J. Accuracy of a smartphone application in estimating sleep in children. Sleep Breath may 2017;21(2):505–11.
[10] Evenson Kelly R, Goto Michelle M, Furberg Robert D. Systematic review of the validity and reliability of consumer-wearable activity trackers. Int J Behav Nutr Phys Activ dec 2015;12(1):159.
[11] Robertson Bonnie, Marshall Buddy, Carno Margaret-Ann. Polysomnography for the sleep technologist. Elsevier Inc.; 2014.
[12] Bani Younis Mohammad, Hayajneh Feryal, Batiha Abdu-Monim. Measurement and nonpharmacologic management of sleep disturbance in the intensive care units. Crit Care Nurs Q jan 2019;42(1):75–80.
[13] Marcus CL, Berry RB, Brooks R, Gamaldo CE, Harding SM, Lloyd RM, Vaughn BV. American Academy of sleep medicine. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications. 2.2 edition. In: American Academy of sleep medicine; 2015.
[14] Dorsch Jennifer J, Martin Jennifer L, Malhotra Atul, Owens Robert L, Kamdar Biren B. Sleep in the intensive care unit: strategies for improvement. Semin Respir Crit Care Med oct 2019;40(5):614–28.
[15] Perez-Pozuelo Ignacio, Zhai Bing, Palotti Joao, Mall Raghvendra,

Table A1
This is the full table of what algorithms where used by the 27 papers chosen for review. It contains the number of sleep stages that were tried to be classified, the type of the algorithms used, the specific algorithm that performed best and their corresponding inputs. When multiple algorithms are mentioned it is because e.g. one was best at 2 stage classification while the other was better in 3 stage classification [29]. does not have an algorithm as they tried many different and did not conclucde which performed best.

Author	No. sleep stages	Alg type	Modality	Algorithm	Input sensor
[38]	2	ML	Multi	SVM	PPG, acc, EDA
[40]	2	DL	Uni	CNN, CNN-LSTM	acc
[51]	2, 5	ML	Uni	RF	acc
[23]	2	ML	Uni	SVM	ECG
[28]	2	ML	Uni	SVM	PPG
[47]	2, 3, 4, 5	DL	Uni, Multi	CNN, LSTM	acc, ECG
[27]	2, 4	ML	Multi	LGBM	PPG, acc, temp
[63]	2	ML	Uni	J48-DT	BCG
[50]	2	DL	Uni	CNN	acc
[31]	2	DL	Uni	LSTM	acc
[39]	3, 4	ML	Multi	LDA	acc, PPG
[37]	3	DL	Uni	CNN-LSTM	BCG
[52]	5	DL	Uni	CUCNN	ECG
[29]	4	ML	Multi		acc, PPG, questionnaire
[33]	4	DL	Multi	MobileNetV2	EEG, EOG
[45]	4	DL	Multi	BLSTM	ECG, acc, abdomon reading
[43]	4	DL	Uni	CNN-BLSTM	BCG
[49]	5	DL	Uni	Time-Distributed CNN	EEG
[44]	2, 5	ML	Uni	Online Stochastic Gradient Descent (OSGD)	acc
[41]	5	ML + DL	Uni	RF, DBM	acc
[48]	5	ML	Uni	SVM	EEG
[34]	4, 5	ML + DL	Uni	RF, DBM	acc
[32]	2, 3, 5	ML	Uni	SVM	In-ear EEG, Scalp-EEG
[42]	2, 3	DL	Multi	MLP	PPG, acc
[46]	4	DL	Uni	MLP	EEG
[30]	2	ML	Multi	DT	PPG, acc, questionnaire
[36]	4	ML	Multi	KNN	Mircowave, infrared

Aupetit Michaël, Garcia-Gomez Juan M, Taheri Shahrad, Guan Yu, Fernandez-Luque Luis. The future of sleep health: a data-driven revolution in sleep science and medicine. npj Digit Med dec 2020;3(1):42.
[16] Ko Ping-Ru T, Kientz Julie A, Choe Eun Kyoung, Kay Matthew, Landis Carol A, Watson Nathaniel F. Consumer sleep Technologies: a review of the landscape. J Clin Sleep Med dec 2015;11(12):1455–61.
[17] Baron Kelly Glazer, Duffecy Jennifer, Berendsen Mark A, Mason Ivy Cheung, Lattie Emily G, Manalo Natalie C. Feeling validated yet? A scoping review of the use of consumer-targeted wearable and mobile technology to measure and improve sleep. Sleep Med Rev aug 2018;40:151–9.
[18] Fallmann Sarah, Chen Liming. Computational sleep behavior analysis: a survey. IEEE Access 2019;7:142421–40.
[19] Fiorillo Luigi, Puiatti Alessandro, Papandrea Michela, Ratti Pietro Luca, Favaro Paolo, Roth Corinne, Bargiotas Panagiotis, Bassetti Claudio L, Faraci Francesca D. Automated sleep scoring: a review of the latest approaches. Sleep Med Rev dec 2019;48:101204.
[20] Janiesch Christian, Zschech Patrick, Heinrich Kai. Machine learning and deep learning. Electron Mark sep 2021;31(3):685–95.
[21] Goodfellow Ian, Bengio Yoshua, Courville Aaron. Deep learning. MIT Press; 2016.
[22] Krizhevsky Alex, Sutskever Ilya, Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. Commun ACM may 2017;60(6):84–90.
[23] Dey Jishnu, Bhowmik Tanmoy, Sahoo Saswata, Tiwari Vijay Narayan. Wearable PPG sensor based alertness scoring system. 2017. In: 2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC); jul 2017. p. 2422–5. IEEE.
[24] Adadi Amina, Berrada Mohammed. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE Access sep 2018;6:52138–60.
[25] Liberati Alessandro, Altman Douglas G, Tetzlaff Jennifer, Mulrow Cynthia, Gøtzsche Peter C, John P, Ioannidis A, Clarke Mike, Devereaux PJ, Kleijnen Jos, Moher David. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med jul 2009;6(7):e1000100.
[26] Vijayan V, Connolly J, Condell J, McKelvey N, Gardiner P. Review of wearable devices and data collection considerations for connected health. Sensors 2021;21(16).
[27] Altini Marco, Kinnunen Hannu. The promise of sleep: a multi-sensor approach for accurate sleep stage detection using the oura ring. Sensors jun 2021;21(13):4302.
[28] Abdul Mohammod, Kamakar MotinxChandan, Marimuthu Palaniswami, Penzel Thomas. Photoplethysmographic-based automated sleep–wake classification using a support vector machine. 7. In: Physiological measurement, vol. 41; aug 2020, 75013.
[29] Liang Zilu, Chapa Martell Mario Alberto. Achieving accurate ubiquitous sleep sensing with consumer wearable activity wristbands using multi-class imbalanced classification. In: 2019 IEEE intl conf on dependable, autonomic and secure computing, intl conf on pervasive intelligence and computing, intl conf on cloud and big data computing, intl conf on cyber science and Technology congress (DASC/PiCom/CBDCom/CyberSciTech); aug 2019. p. 768–75. IEEE.
[30] Liang Zilu, Chapa-Martell Mario Alberto. Combining resampling and machine learning to improve sleep-wake detection of fitbit wristbands. In: 2019 IEEE international conference on healthcare informatics (ICHI); jun 2019. p. 1–3. IEEE.
[31] Yildiz Selda, Opel Ryan A, Elliott Jonathan E, Kaye Jeffrey, Cao Hung, Lim Miranda M. Categorizing sleep in older adults with wireless activity monitors using LSTM neural networks. In: 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC); jul 2019. p. 3368–72. IEEE.
[32] Nakamura Takashi, Alqurashi Yousef D, Morrell Mary J, Hearables Danilo P Mandic. Automatic overnight sleep monitoring with standardized in-ear EEG sensor. IEEE (Inst Electr Electron Eng) Trans Biomed Eng jan 2020;67(1):203–12.
[33] Tsung Hao Hsieh T-H, Liu Meng Hsuan M-H, Chin En Kuo C-E, Wang Yung Hung Y-H, Sheng Fu Liang S-F. Home-use and real-time sleep-staging system based on eye masks and mobile devices with a deep learning model. In: Journal of medical and biological engineering; sep 2021. p. 1–10.
[34] Reimer Ulrich, Emmenegger Sandro, Maier Edith, Ulmer Tom, Vollbrecht Hans-Joachim, Zhang Zhongxing, Khatami Ramin. Laying the foundation for correlating daytime behaviour with sleep architecture using wearable sensors. In: Molloy W, Rocker C, Ziefle M, Maciaszek LO, Donoghue J, editors. Communications in computer and information science, vol. 869; apr 2018. p. 147–67. Springer Verlag.
[35] Wang Chiapin, Chiang Tsung-Yi Fan, Fang Shih-Hau, Li Chieh-Ju, Hsu Yeh-Liang. Machine learning based sleep-status discrimination using a motion sensing mattress. In: 2019 IEEE international conference on artificial intelligence circuits and systems (AICAS); mar 2019. p. 160–2. IEEE.
[36] Yoon Yeong Sook, Hahm Jarang, Kim Kwang Ki, Park Su Kyung, Oh Sang Woo. Non-contact home-adapted device estimates sleep stages in middle-aged men: a preliminary study. Technol Health Care jul 2020;28(4):439–46.
[37] Gargees Rayan, Keller James M, Popescu Mihail, Skubic Marjorie. Non-invasive classification of sleep stages with a hydraulic bed sensor using deep learning. 11862 LNCS. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics); oct 2019. p. 73–82. Springer, Cham.
[38] Assaf Mahmoud, Rizzotti-Kaddouri Aïcha, Punceva Magdalena. Sleep detection using physiological signals from a wearable device. In: EAI/Springer innovations in communication and computing; nov 2020. p. 23–37. Springer Science and Business Media Deutschland GmbH.
[39] Fedorin Illia, Slyusarenko Kostyantyn, Lee Wonkyu, Sakhnenko Nataliya. Sleep stages classification in a healthy people based on optical plethysmography and accelerometer signals via wearable devices. In: 2019 IEEE 2nd Ukraine conference on electrical and computer engineering (UKRCON); jul 2019. p. 1201–4. IEEE.
[40] Cho Taeheum, Sunarya Unang, Yeo Minsoo, Hwang Bosun, Seo Koo Yong, Park Cheolsoo. Deep-ACTINet: end-to-end deep learning architecture for automatic sleep-wake detection using wrist actigraphy. Electronics dec 2019;8(12):1461.
[41] Reimer Ulrich, Emmenegger Sandro, Maier Edith, Zhang Zhongxing, Khatami Ramin. Recognizing sleep stages with wearable sensors in everyday settings. In: Proceedings of the 3rd international conference on information and communication Technologies for ageing well and e-health, number january, pages 172–179; 2017. SCITEPRESS - Science and Technology Publications.
[42] Walch Olivia, Huang Yitong, Forger Daniel, Goldstein Cathy. Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. Sleep dec 2019;42(12):1–19.
[43] Rao Shashank, El Ali Abdallah, Cesar Pablo. DeepSleep: a ballistocardiographic deep learning approach for classifying sleep stages. New York, NY, USA. In: Adjunct proceedings of the 2019 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2019 ACM international symposium on wearable computers; sep 2019. p. 187–90. ACM.
[44] Sajjad Hossain HM, Ramamurthy Sreenivasan R, Hafiz Khan Abdullah Al, Roy Nirmalya. An active sleep monitoring framework using wearables. ACM Trans. Interact. Intel. Syst aug 2018;8(3):1–30.
[45] Zhang Yuezhou, Yang Zhicheng, Lan Ke, Liu Xiaoli, Zhang Zhengbo, Li Peiyao, Cao Desen, Zheng Jiewen, Pan Jianli. Sleep stage classification using bidirectional LSTM in wearable multi-sensor systems. In: I. EEE INFOCOM 2019 - IEEE conference on computer communications workshops (INFOCOM WKSHPS); apr 2019. p. 443–8. IEEE.
[46] Zhang Zhuo, Guan Cuntai. An accurate sleep staging system with novel feature generation and auto-mapping. In: 2017 international conference on orange Technologies (ICOT), volume 2018-janua; dec 2017. p. 214–7. IEEE.
[47] Zhai Bing, Perez-Pozuelo Ignacio, Clifton Emma AD, Palotti Joao, Guan Yu. Making sense of sleep. In: Proceedings of the ACM on interactive, mobile, wearable and ubiquitous Technologies, vol. 4; jun 2020. p. 1–33. 2.
[48] Ali Zamin Syed, Awais Bin Altaf Muhammad, Saadeh Wala. A single channel EEG-based all AASM sleep stages classifier for neurodegenerative disorder. IEEE. In: 2019 IEEE biomedical circuits and systems conference (BioCAS); oct 2019. p. 1–4.
[49] Koushik Abhay, Judith Amores, Maes Pattie. Real-time smartphone-based sleep staging using 1-channel EEG. In: 2019 IEEE 16th international conference on wearable and implantable body sensor networks (BSN); may 2019. p. 1–4. IEEE.
[50] Peraza Luis R, Joules Richard, Dauvilliers Yves, Wolz Robin. Device agnostic sleep-wake segment classification from wrist-worn accelerometry. In: 2020 IEEE international conference on healthcare informatics (ICHI); nov 2020. p. 1–3. IEEE.
[51] Sundararajan Kalaivani, Georgievska Sonja, Bart H, te Lindert WW, Gehrman Philip R, Ramautar Jennifer, Mazzotti Diego R, Sabia Séverine, Weedon Michael N, Eus J, W van Someren W, Ridder Lars, Wang Jian, Vincent T, van Hees. Sleep classification from wrist-worn accelerometer data using random forests. Sci Rep dec 2021;11(1):24.
[52] Zhang Junming, Wu Yan. Complex-valued unsupervised convolutional neural networks for sleep stage classification. Comput Methods Progr Biomed oct 2018;164:181–91.
[53] Borazio Marko, Berlin Eugen, Kucukyildiz Nagihan, Scholl Philipp, Van Laerhoven Kristof. Towards benchmarked sleep detection with wrist-worn sensing units. In: 2014 IEEE international conference on healthcare informatics; sep 2014. p. 125–34. IEEE.
[54] Chen Xiaoli, Wang Rui, Zee Phyllis, Lutsey Pamela L, Javaheri Sogol, Alcántara Carmela, Jackson Chandra L, Williams Michelle A, Redline Susan. Racial/ethnic differences in sleep disturbances: the Multi-Ethnic Study of Atherosclerosis (MESA). Sleep jun 2015;38(6):877–88.
[55] Zhang Guo Qiang, Cui Licong, Mueller Remo, Tao Shiqiang, Kim Matthew, Rueschman Michael, Mariani Sara, Mobley Daniel, Redline Susan. The national sleep research resource: towards a sleep data commons. J Am Med Inf Assoc oct 2018;25(10):1351–8.
[56] Moody GB, Mark RG. The impact of the MIT-BIH arrhythmia database. IEEE Eng Med Biol Mag 2001;20(3):45–50.
[57] Montgomery-Downs Hawley E, Insana Salvatore P, Bond Jonathan A. Movement toward a novel activity monitoring device. Sleep Breath sep 2012;16(3):913–7.
[58] 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 dec 1997;20(12):1077–85.
[59] Kemp B, Zwinderman AH, Tuk B, Kamphuisen HAC, Oberye JJL. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. IEEE (Inst Electr Electron Eng) Trans Biomed Eng 2000;47(9):

1185–94.
[60] Giovanni Terzano Mario, Parrino Liborio, Sherieri Adriano, Chervin Ronald, Chokroverty Sudhansu, Guilleminault Christian, Hirshkowitz Max, Mahowald Mark, Harvey Moldofsky, Rosa Agostino, Thomas Robert, Walters Arthur. Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep. Sleep Med nov 2001;2(6):537–53.
[61] Vincent van Hees, Charman Sarah, Anderson Kirstie. Newcastle polysomnography and accelerometer data. January. 2018.
[62] Shrivastava Deepak, Jung Syung, Saadat Mohsen, Sirohi Roopa, Crewson Keri. How to interpret the results of a sleep study. J Community Hosp Intern Med Perspect jan 2014;4(5):24983.
[63] 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. 2019 PeerJ 2019;(9).
[64] den van Oord Aaron, Sander Dieleman, Zen Heiga, Simonyan Karen, Vinyals Oriol, Graves Alex, Kalchbrenner Nal, Senior Andrew, Kavukcuoglu Koray. WaveNet: a generative model for raw audio. 2016.
[65] Wilkinson Mark D, Dumontier Michel, Aalbersberg IJsbrand Jan, Appleton Gabrielle, Axton Myles, Baak Arie, Blomberg Niklas, Boiten Jan-Willem, Silva Santos Luiz Bonino da, Bourne Philip E, Bouwman Jildau, Brookes Anthony J, Clark Tim, Crosas Mercè, Dillo Ingrid, Olivier Dumon, Scott Edmunds, Evelo Chris T, Finkers Richard, Gonzalez-Beltran Alejandra, Alasdair J, Gray G, Groth Paul, Goble Carole, Grethe Jeffrey S, Heringa Jaap, Hoen Peter AC 't, Hooft Rob, Kuhn Tobias, Kok Ruben, Kok Joost, Lusher Scott J, Martone Maryann E, Albert Mons, Packer Abel L, Persson Bengt, Rocca-Serra Philippe, Roos Marco, van Schaik Rene, Sansone Susanna-Assunta, Erik Schultes, Sengstag Thierry, Slater Ted, George Strawn, Swertz Morris A, Thompson Mark, van der Lei Johan, van Mulligen Erik, Jan Velterop, Waagmeester Andra, Peter Wittenburg, Wolstencroft Katherine, Zhao Jun, Mons Barend. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data dec 2016;3(1):160018.
[66] Faust Oliver, Razaghi Hajar, Barika Ragab, Ciaccio Edward J, Acharya U Rajendra. A review of automated sleep stage scoring based on physiological signals for the new millennia. jul 2019.
[67] Radha Mustafa, Fonseca Pedro, Moreau Arnaud, Ross Marco, Cerny Andreas, Anderer Peter, Long Xi, Ronald M, Aarts. Sleep stage classification from heart-rate variability using long short-term memory neural networks. Sci Rep dec 2019;9(1):14149.
[68] Radha Mustafa, Fonseca Pedro, Moreau Arnaud, Ross Marco, Cerny Andreas, Anderer Peter, Long Xi, Aarts Ronald M. A deep transfer learning approach for wearable sleep stage classification with photoplethysmography. npj Digit Med dec 2021;4(1):135.
[69] Zhu Qiang, Tian Xin, Wong Chau-Wai, Wu Min. ECG reconstruction via PPG: a pilot study. IEEE. In: 2019 IEEE EMBS international conference on biomedical & health informatics (BHI); may 2019. p. 1–4.
[70] Zemblys Raimondas, Diederick C. Niehorster, and Kenneth Holmqvist. gazeNet: end-to-end eye-movement event detection with deep neural networks. Behav Res Methods oct 2018:1–25.
[71] Ryan Connaughton, Bowyer Kevin W, Flynn Patrick J. Chapter 17 fusion of face and Iris biometrics. In: Handbook of Iris recognition; 2016. p. 397–415.
[72] Eitel Andreas, Springenberg Jost Tobias, Spinello Luciano, Riedmiller Martin, Burgard Wolfram. Multimodal deep learning for robust RGB-D object recognition. 2015-Decem. In: IEEE international conference on intelligent robots and systems; jul 2015. p. 681–7.
[73] Fahimi Fatemeh, Zhang Zhuo, Goh Wooi Boon, Lee Tih-Shi, Ang Kai Keng, Guan Cuntai. Inter-subject transfer learning with an end-to-end deep convolutional neural network for EEG-based BCI. J Neural Eng apr 2019;16(2):026007.
[74] van Berkel Niels, Skov Mikael B, Kjeldskov Jesper. Human-AI interaction. Interactions nov 2021;28(6):67–71.
[75] Janlert Lars-Erik, Stolterman Erik. The meaning of interactivity—some proposals for definitions and measures. Hum Comput Interact may 2017;32(3):103–38.
[76] Yang Qian, Steinfeld Aaron, Rosé Carolyn, Zimmerman John. Re-Examining whether, why, and how human-AI interaction is uniquely difficult to design. New York, NY, USA. In: Proceedings of the 2020 CHI conference on human factors in computing systems; apr 2020. p. 1–13. ACM.
[77] Amershi Saleema, Dan Weld, Vorvoreanu Mihaela, Adam Fourney, Nushi Besmira, Collisson Penny, Suh Jina, Iqbal Shamsi, Bennett Paul N, Inkpen Kori, Jaime Teevan, Kikin-Gil Ruth, Horvitz Eric. Guidelines for human-AI interaction. In: Proceedings of the 2019 CHI conference on human factors in computing systems. New York, NY, USA: ACM; may 2019. p. 1–13.