Sleep Medicine Reviews 48 (2019) 101204

Contents lists available at ScienceDirect

Sleep Medicine Reviews

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

CLINICAL REVIEW

Automated sleep scoring: A review of the latest approaches

Luigi Fiorillo [a, c], Alessandro Puiatti [a], Michela Papandrea [a], Pietro-Luca Ratti [b], Paolo Favaro [c], Corinne Roth [d], Panagiotis Bargiotas [d, e], Claudio L. Bassetti [d], Francesca D. Faraci [a, *]

[a] Institute for Information Systems and Networking, University of Applied Sciences and Arts of Southern Switzerland, Manno, Switzerland
[b] Clinical Neurophysiology Unit, Department of Neurology, Pierre Zobda-Quitman Hospital, University Hospitals of Martinique, Fort-de-France, Martinique, France
[c] Institute of Informatics, University of Bern, Bern, Switzerland
[d] Department of Neurology, Sleep-Wake-Epilepsy Center, Inselspital University Hospital Bern, University of Bern, Bern, Switzerland
[e] Department of Neurology, Medical School, University of Cyprus, Nicosia, Cyprus

ARTICLE INFO

Article history:
Received 4 February 2019
Received in revised form
11 July 2019
Accepted 22 July 2019
Available online 9 August 2019

Keywords:
Sleep scoring
Automated and semi-automated systems
Artificial intelligence
Shallow learning
Deep learning

SUMMARY

Clinical sleep scoring involves a tedious visual review of overnight polysomnograms by a human expert, according to official standards. It could appear then a suitable task for modern artificial intelligence algorithms. Indeed, machine learning algorithms have been applied to sleep scoring for many years. As a result, several software products offer nowadays automated or semi-automated scoring services. However, the vast majority of the sleep physicians do not use them. Very recently, thanks to the increased computational power, deep learning has also been employed with promising results. Machine learning algorithms can undoubtedly reach a high accuracy in specific situations, but there are many difficulties in their introduction in the daily routine. In this review, the latest approaches that are applying deep learning for facilitating and accelerating sleep scoring are thoroughly analyzed and compared with the state of the art methods. Then the obstacles in introducing automated sleep scoring in the clinical practice are examined. Deep learning algorithm capabilities of learning from a highly heterogeneous dataset, in terms both of human data and of scorers, are very promising and should be further investigated.

Introduction

Sleep disorders represent a significant and increasing public health problem; a considerable proportion of the world population is suffering from serious sleep disorders and is requiring medical attention [1]. Whole night polysomnography (PSG), originated in the late 1950s, is the gold standard to evaluate sleep and to identify sleep disorders.

The PSG monitors brain activity (EEG), eye movements (EOG), muscle activity or skeletal muscle activation (EMG derivations for chin and legs), body position (video camera) and heart rhythm (ECG). Breathing functions (respiratory airflow, oxygen saturation, respiratory effort indicators) are also measured. A PSG typically requires that the patients sleep overnight at the hospital while their bio-physiological signals are recorded.

Sleep scoring is the process of extracting sleep cycle information from the electrophysiological signals. Sleep stages, arousals, respiratory events, movements and cardiac events have to be correctly identified. Wakefulness and sleep phases are described by three bio-signals: EEG, EOG and EMG. To score an eight-hour PSG may require up to two hours of tedious repetitive work. This can explain why the search for simplifying and speeding up the sleep scoring work begun already in the late 1960s. In fact, given the characteristics of the sleep scoring, it appeared to be a classical problem to be solved in an automated approach using an information processing system. Many different techniques and approaches have been proposed and tested, reaching very good results in terms of accuracy. Automatic sleep scoring basic principles have been accurately presented by Penzel et al. [2].

However, automatic scoring is not yet an everyday reality in sleep centers worldwide. Certainly, as reported by [3–5], the high inter-scorer variability (agreement of about 70–80%), and the low intra-scorer agreement (about 90%), reduce the acceptance of

* Corresponding author. University of Applied Sciences and Arts of Southern Switzerland, Via Cantonale 2c, 6928, Manno, Switzerland.
E-mail address: francesca.faraci@supsi.ch (F.D. Faraci).

https://doi.org/10.1016/j.smrv.2019.07.007
1087-0792/

Abbreviations

AASM	American Academy of Sleep Medicine
AI	artificial intelligence
ANN	artificial neural network
ASSC	automatic sleep stage classification
CNN	convolutional neural network
DBN	deep belief network
DNN	deep neural network
DWT	discrete wavelet transform
EEG	electroencephalography
ECG	electrocardiography
EDF	European data format
EMG	electromyography
EOG	electrooculogram
LSTM	long short term memory
ML	machine learning
MT	movement time
PCA	principal component analysis
PSG	polysomnography
R&K	Rechtschaffen and Kales
REM	rapid eye movement
RF	random forest
RNN	recurrent neural network
RUS-Boost	random under-sampling boosting
SFSM	sequential floating search method
SVM	support vector machine

automated scoring systems. A systematic review, detailed in the methodology section, of the application of emerging deep learning algorithms to sleep scoring is here presented. The aim is to understand if the adoption of automatic scoring in the clinical practice can be finally facilitated thanks to this new approach. Firstly, visual scoring procedure and its complexity is presented. Then, after a short presentation of the general artificial intelligence (AI) algorithm application in sleep scoring, existing deep learning techniques are thoroughly examined. Alternative methodologies like EOG single-channel systems and semi-automated procedures, are then presented. For completeness commercially available software are also summarized. In the last section, the general barriers to the introduction of automatic scoring in the clinical practice are presented, trying to understand if a deep learning based methodology can overcome at least some of these barriers.

Visual scoring procedure

In order to understand automated scoring problematics there is need first to analyze what AI is trying to reproduce: the visual scoring process. The polysomnographic record of sleep is usually divided into 30-second epochs, starting from the lights-off event. This time interval is a heritage from old PSG machines where a paper speed of 10 mm/s gave an output page of a 30-second timespan. During a visual analysis each epoch is assigned a stage, and if two or more stages coexist during a single epoch the stage comprising the majority of the 30 seconds is scored.

In 1968 the first manual to standardize terminology and rules of this procedure was published by Rechtschaffen and Kales (R&K) [6]. It categorized sleep into seven distinct stages: wakefulness, stages 1, 2, 3, 4, rapid eye movement (REM) sleep and movement time (MT) stage. These rules were adopted worldwide until 2007, when the American Academy of Sleep Medicine (AASM) updated the scoring manual [7]. The AASM standard manual for the scoring of sleep and associated events is designed to cover all aspects of the PSG, from the technical ones (parameters, assessment protocols, filtering, etc.), to its execution, the analytic scoring (sleep staging, arousals, cardiac, movement, and respiratory signals), and the final interpretation of PSG results. The number of stages was reduced to five: wakefulness W, stage N1, stage N2, stage N3 (formerly stages 3 and 4 sleep), and stage R sleep (formerly stage REM sleep). MT stage was abolished, and it was decided to score an epoch with a major body movement as wake if any part of the epoch shows alpha rhythm, or if a wake epoch either precedes or follows the epoch in question. Otherwise, the epoch is scored as the same stage as the epoch that follows it. Almost every year, to date 2018 (v2.5), there is a new version of the AASM manual with usually a few updates.

Recommended EEG derivations are F4-M1, C4-M1, O2-M1, while other accepted derivations are Fz-Cz, Cz-Oz, C4-M1. EEG can be contaminated by other electrophysiological signals, as for example ECG, EOG, EMG and pulse-oxymetry signal. Movement artifacts are also often present and need to be addressed. EEG is conventionally described in terms of its frequency components. The main ones are delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), and beta (12–35 Hz). Waves in the frequency range 0.5–2 Hz and peak-to-peak amplitude >75 μV are considered slow wave activity. Sleep spindles (train of distinct waves in the 12–14 Hz range, lasting for more than 0.5s), K-complexes (sharp negative waves followed by a smooth, positive waves longer than 0.5s) and vertex sharp waves (negative-going bursts of less than 0.5s) are also introduced to better describe the EEG. Scoring rules are based on the recognition of EEG frequencies and on the presence of certain pattern, but applying these rules can lead to unexpected complexity, especially in unhealthy subjects.

Sleep stages progress cyclically from N1 through R, then begin again with stage N1. A full sleep cycle takes on average from 90 to 110 min, with each stage lasting between 5 and 15 min. The first sleep cycles present relatively short REM sleeps and long periods of deep sleep. The characteristics of the sleep phases and the scoring rules accordingly to the AASM are summarized in the following list.

- ***Stage W***: it is characterized by the presence of alpha rhythm in the EEG signal, usually over the occipital region, and/or any of the following events: eye blinking, rapid eye movements with normal or high chin muscle activity (with signal frequency higher than 30 Hz), reading eye movements.
- ***Stage N1***: it shows slow eye movements and it can be easily disrupted leading to awakenings or arousals. EEG signal amplitude does not exceed 200 mV with frequencies within 2–7 Hz. Alpha components should not exceed 50% of the total spectral band, and vertex sharp waves are often seen during transitions from other stages to N1. Slow eye movements can be visible in the EOG, and EMG level should be lower than in the previous stage. N1 continues until there is evidence of another stage. N1 usually covers 5% of the total sleep time.
- ***Stage N2***: awakenings or arousals are not so common as in N1 and the slow-moving eye starts to disappear. Sleep spindles and K complexes may appear. N2 should be scored if during the last half of the previous epoch or the first half of the actual one there are either one or more K-complex or one or more trains of spindle, and it should continue to be scored N2 (also without spindles and K-complexes), until a new stage appear. N2 normally covers 50% of the total sleep time.

- ***Stage N3***: it is the deep restorative sleep. Delta waves and slow waves are predominant in the EEG signal. Awakenings or arousals are rare. Spindles and K-complex may appear. N3 should be scored if more than 20% of the epoch consists of slow waves. N3 covers usually 20% of the total sleep time.
- ***Stage REM***: the dreaming stage. Eye movements are rapid and brain waves are more active than in N2 and in N3. Awakenings and arousals can occur more easily in REM. The EEG has low voltage, mixed frequency and possible sawtooth waves, EMG is at its lowest level, episodic REMs usually lasting less than 500 ms appears in the EOG. A stage should continue to be scored as REM until one or more of the following occur: a transition to stage W or N3 appears; chin EMG muscle tone increases; a K-complex without arousal or a spindle occurs in the first half of the epoch with no REMs. This stage normally covers 20–25% of the total sleep time.

The sleep staging procedure may be quite complex: many parameters have to be considered at the same time; previous and future epoch scoring has to be taken into account as well. The heterogeneity of the subjects and of the sleep epochs may be quite difficult to be comprehensively described in a manual, which generates uncertainty in the scoring procedure and leads to different interpretations of the same signal from different scorers. Recording from subjects with specific sleep disorders can be more challenging to be scored than healthy ones. It has also to be highlighted that some errors may be more costly than others. N2 is very often considered as a transition phase between light and deep sleep, consequently if N2 percentage is a little higher or a little lower, the impact on the "big picture" of the sleep analysis will be minimal. The presence of sleep apneas, parasomnias, periodic limb movement and of other sleep abnormalities will still be examined if N2 is confused with N1 or N3. Instead, if the error is related to wakefulness all the scoring process will be impacted, as the presence of sleep abnormalities will not be considered. Inter-rater variability studies show how agreement can vary among stages [8]. Rosemberg et al. [3] compared a large number of scorers (>2500): the agreement was higher than 80% for REM, N2 and W, but it dropped for N3 (67%) and N1 (63%), the overall agreement was of about 83%. Human scorers' discrepancies occur mainly in the judgment of transitions between two different stages. This is not surprisingly as AASM rules are trying to characterize a continuum physiological process with fixed stages.

Methodology

A systematic review of the recent application of deep learning techniques to sleep scoring was conducted and is here reported accordingly to the PRISMA statement guidelines [9].

- ***Search strategy***: electronic searches in PubMed and Web of Science databases were conducted to identify all relevant studies published between 2016 and March 2019. The following keywords, selected also from the MeSH database, have been considered: (*"deep learning" [All fields] OR "deep neural network*" [All fields] OR "convolutional neural network*" [All fields] OR "recurrent neural network*" [All fields] OR "CNN*" [All fields] OR "RNN*" [All fields] OR "LSTM*" [All fields]) AND ("sleep" [All fields] OR "sleep scoring" [All fields] OR "sleep stage*" [All fields] OR "sleep staging" [All fields])*.
- ***Identification***: the total number of studies identified was 135, of which 48 records in PubMed and 87 records in Web of Science. The bibliography/reference lists of included studies were also reviewed, three more studies were then identified, reaching a number of 138 records.
- ***Screening***: Applying the inclusion criteria and removing duplicates between the two databases 27 records were selected by the first and the last authors, in case of doubt the other authors were consulted. The most recent contribution by the same research group was selected.
- ***Inclusion criteria***: studies were included if they met the following criteria: 1) PSG recordings belonging to adults (above 18 y old); 2) automatic sleep scoring algorithms using EEG signals only or in combination with EOG and/or EMG signals; 3) scoring according to the R&K or to the AASM rules in five sleep states (stages 3 and 4 sleep considered as a single stage N3 or slow wave sleep); 4) PSG data recorded in a clinical environment or in a sleep laboratory; 5) deep learning algorithm applied directly to raw signals or to spectrograms images and 6) English language and peer-reviewed journal, conferences or workshop. Only one non-peer-reviewed publication was included, as it was referenced by many peer-reviewed publications.
- ***Data extraction***: 14 studies were finally included in the review. For each study the following information were extracted: dataset, subject type (healthy or not), number of PSG channels and derivations, classifier type, performance reached.
- ***Synthesis of results***: The heterogeneity of the studies in terms of methodologies and datasets employed makes a meta-analysis inappropriate, thus a systematic qualitative review of the literature was conducted.

Automatic sleep scoring by artificial intelligence

Artificial intelligence consists of the emulation of human intelligence processes performed by machines. Machine learning (ML) is an application of AI that provides systems the ability to automatically learn and improve from experience. ML algorithms can then replicate human intelligence processes to assist and simplify manual procedure. This powerful approach should then be suitable for sleep scoring, which could be considered as a tedious, repetitive work based on the observation of standardized rules.

Two main approaches of AI might potentially address the automatic sleep stage classification (ASSC) problem: learning processes based on features extracted starting from the knowledge of the experts (shallow learning), and learning processes that start directly from the raw data (deep learning).

Shallow learning – expert knowledge

In a machine learning workflow, the main steps are data preprocessing, feature extraction, feature selection/dimensionality reduction and classification. The pre-processing phase allows the detection of bias, noise or artifact present in the PSG raw signals. The feature extraction, feature selection and dimensionality reduction steps allow to identify the most relevant information. The last classification phase uses all the information for sleep stage identification.

Feature extraction and feature selection techniques

The feature extraction procedure starts from the measured data and derives values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to a better human interpretation. A feature is an individual measurable property or characteristic of the PSG; an example can be the time-domain signal power over the entire epoch. Feature extraction techniques can be linear and non-linear and can be grouped into three major categories: temporal domain methods, frequency domain methods and hybrid of temporal and frequency domain methods [10,11].

Among the most recent works, the standard statistics in time domain, the non-parametric analysis in frequency domain and the wavelet transform in time-frequency domain are the most used techniques for ASSC [12].

In some cases the extracted features are redundant or generate a dataset with a high dimensionality. Dimensionality reduction is a process that can reduce the number of features, focusing on the most significant ones. The most widely used approaches are the principal component analysis (PCA), for dimensionality reduction, and the sequential floating search method (SFSM) for feature selection [12]. These techniques allow representing data in a reduced dimensional space maintaining almost the same information, resulting in increased performance of the classifier [13].

Machine learning classifiers

The machine learning classification techniques used in automatic sleep stage identification – shallow learning approach – are manifold. Several reviews have exhaustively analyzed feature-based approaches. In particular, Ronzhina et al. [14] reviewed classification systems using artificial neural networks (ANN) in automatic sleep scoring. The reported ANN-based scoring system performance varies within a broad range of accuracy, depending on the recognized stages. Şen et al. [11] carried out a comparative study trying to identify the most effective features and the most efficient algorithm to classify the sleep stages. They propose a methodology that can reach an overall accuracy of 98%. Radha et al. [15] also tried to identify optimal ML and signal processing methods, focusing on online sleep staging and a single EEG channel. They concluded that spectral linear features, epoch duration between 18 and 30 s, and a random forest (RF) classifier lead to optimal classification performance while ensuring real-time online operation. In the comprehensive survey of Aboalayon et al. [12] several sleep stage classification techniques using EEG signals have been reported and compared, with accuracy ranging from 70 to 94%. They have also presented their own approach based on novel features and using 10-second epochs claiming to reach an average accuracy of 93%.

It is important to note that comparing the performance of different approaches is a quite complex task. Sleep stages considered, extracted features, datasets and channels, classification algorithms, validation methods adopted and evaluation metrics reported have to be taken into consideration. For a better comparison, some researchers have reapplied classification approaches to the same dataset. For example, in a recent work, Boostani et al. [16] carried out a comparative review of several machine learning classification techniques used in ASSC. They selected five classification approaches [17–21] and reapplied them on public datasets containing PSG data of healthy and unhealthy subjects. They tested various combinations of extracted features and classification techniques in order to find the best one in predicting sleep stages correctly. The random forest classifier together with the entropy of wavelet coefficients proved to be the best combination, reporting percentages of accuracy of 87% in healthy subjects and 69% in patients.

Deep learning algorithms can also be applied in the feature-based workflow with good results [22,23], but they exert their full potential when applied directly to raw data, as presented in the next paragraph.

Deep learning – knowledge from raw data

Deep learning is part of a broader family of machine learning methods; it is based on learning data representations, as opposed to task-specific algorithms. In recent years the use of deep learning classification techniques has shown to be highly performing in several fields of application such as image captioning, image classification, and speech recognition [24–26]. The possibility to extract complex information from a large amount of data is one of the first reasons to apply deep learning techniques in PSG classification.

The great advantage of the deep learning models is the high performance in dealing with a large amount of data. DL can learn features directly from the raw input data with little to no prior knowledge. However, the non-interpretability of the results and the longer computational times can be a drawback. On the other hand, features extracted starting from the knowledge of the experts are thought to be affected by several factors, primarily by the characteristics of the available dataset [27]. In sleep scoring the dataset present a wide variety, and the number of epochs in a single dataset is huge. The feature-based approach may not be suitable to satisfy a comprehensive description of the heterogeneity of the subjects and the set of recorded signals. For this reason, over the last few years, several works applied deep learning algorithms directly on raw PSG signals.

CNN and RNN architectures

Deep learning models are based on artificial neural networks and differ mainly on the architecture, which is how several neurons are arranged and connected to each other. Neurons lying on the same level make up the so-called *layer*. The network is composed of several units or neurons, each of them performs a linear combination of the input followed by a non linear transformation, as explained in details in Fig. 1(a).

The standard deep neural networks are characterized by multiple layers sequentially fully connected. Several types of deep learning architectures have been developed. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are the most widely used in ASSC.

A convolutional neural network is a supervised classification model in which the input (e.g., raw data, spectrogram images) is processed by a network of filters and sub-sampling (pooling) layers. Each of these filters can be thought of as feature identifiers, whereas sub-sampling reduces the dimensionality but retains the important information. The last layer, usually a softmax layer, computes the output probability of each sleep stage to identify the target of the signal. An example of convolutional neural network overall architecture is provided in Fig. 1(b). During the training phase of the CNN the neuron filter weights and bias are adjusted in order to reach the target probability class for that input (epochs). After training, the CNN is ready to be applied on new input. As more layers are stacked more complex features are produced. Also the RNNs are networks of filters that can be trained, but they work on the principle of saving the output of a layer and feeding it back to the input, in order to predict the future output of the layer. CNN considers only the current input while RNN considers the current input together with the previously received inputs. Therefore, conversely from CNN, RNNs can handle sequential data.

In a sleep stage scoring procedure, the staging of each 30-second epoch is strongly related to the preceding and following epochs. Thus, a temporal dynamic behavior unit has to be added to the CNN by introducing recurrent connections. The more common memory units are a type of RNN called long-short-term memory (LSTM). This memory allows the model to process sequences of inputs (epochs in the sleep staging case).

Evaluation of deep learning in sleep scoring

In compliance with the criteria defined in the methodology section, 14 publications have been included for analyzing recent deep learning techniques applied to sleep scoring. The following

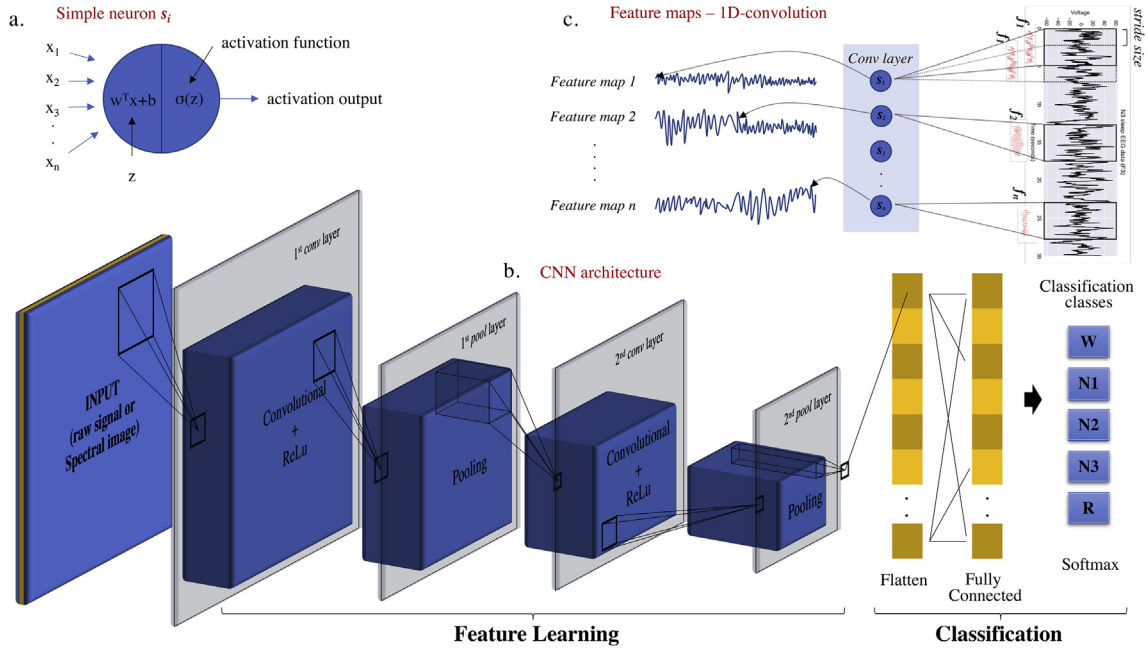

Fig. 1. CNN architectures and sleep staging. (a) Each single neuron computes the dot product of the input ***x*** and the weight ***w*** vectors. A bias ***b*** value is added to the dot product and a non-linear function, called activation function (e.g., sigmoid σ, hyperbolic tangent *tanh*, rectified linear unit or *ReLu*, leaky *ReLu*), is then applied. (b) The CNN architecture can be divided in two subsequent parts, each performing a different process. The first, the feature learning activity, consists of several convolutional *conv* and of some pooling *pool* layers. The last, the classification process, is carried out by a fully connected layer and a softmax function. (c) Example of the 1D-convolution operation and feature maps construction in a *conv* layer. *Conv* layer firstly implements the convolution operation between the 30-second epoch and the n filters ***f*i** of the n neurons ***s*i**, then an activation function is applied (e.g., *ReLu*). Each feature map *i* is the output of each convolution operation. Each value of a feature map can be considered as the result of the dot product between the local part of the input (size of the filter) and the filter ***f***. The dash-line window shows how each filter ***f*i** is shifted during the convolution (stride). The signal is flattened to one dimension, it is processed through a fully connected layer and finally classified using the last softmax layer. Unlike the *conv* layer, every unit of the fully connected layer interact with every input unit. The softmax layer computes the output probability of each sleep stage to identify the target of the signal.

information has been extracted: dataset characteristics, subject type (healthy or unhealthy), information sources (channels), deep learning network type (classifier) and performance. A summary is given in Table 1.

Dataset characteristics

Several public and not public PSG datasets are employed. The more common are Sleep-EDF [28] and Sleep-EDF[Expanded] [29], followed by the Montreal archive of sleep studies [30] and the sleep heart health study collection [31]. The biggest dataset belongs to the Massachusetts General Hospital Sleep Laboratory, which contains 10,000 recordings. One night recording has around 800 epochs, so also a small dataset may allow the application of deep learning algorithms. In fact, some authors train their models with less than 30 recordings [32,33]. Datasets may differ for the sampling rate and for the hardware pre-processing. Sometimes the subject category (healthy or unhealthy) as well as the human scorer identifier is missing.

As explained in the visual scoring procedure section, in one night sleep W and N1 epochs are less frequent than others. As a consequence, PSG datasets are not balanced with respect to the number of classification targets (sleep stages). Usually there are a lot more epochs for N2 stage than for W and N1 stages. Without balancing the dataset, it is highly likely that a classifier will exhibit skewed performance favoring the most represented classes, unless the least represented are very distinct from the other ones. *Balanced sampling* is used to overcome this issue. There exist different possibilities: some authors apply *oversampling* on the sleep stages with fewer samples [27]; others apply *under-sampling* on the sleep stages with a higher number of samples [34,35]. In some other studies, a weight is computed for each class, defined as the ratio between the frequency of the single class divided by the frequency of the most frequent class. The weights are assigned to each class as to contribute equally to the final prediction [36].

Subject type

The datasets contain usually data from both healthy and unhealthy subjects. The work of Patanaik et al. [37] shows that training the classifier on healthy subjects and then validating it on patients leads to a lower agreement. As it could have been expected, the lowest agreement is usually with patients with neurodegenerative diseases, such as Parkinson's. Indeed, the loss of structure or function of neurons leads to important alteration of the and to the destructuring of sleep architecture.

Information sources — channels

The EEG channel derivations employed in the PSG recordings in different datasets are both the ASSM recommended and the acceptable ones. Several studies exploited the information from different EEG channels, often combined with EMG and EOG signals. Chambon et al. [34] and Cui et al. [38] propose the works that exploit the highest variety of information. They utilize in fact six EEG channels with also EMG and EOG. Other authors used two EEG or six EEG channels, with or without EMG and/or EOG signals. Around half of the works are EEG single-channel based. Generally, it has been shown that the performance increases, although not always significantly, with a multi-channel EEG approach.

Table 1
Summary of systems for sleep scoring using deep learning classification techniques directly applied to rawdata. We report: the dataset from which the data were extracted; the type and number of subjects considered in the analysis; the type and number of channels taken into account; the type of deep learning classification algorithms and the best performance achieved. ANN: artificial neural network, CNN: convolutional neural network, EEG: electroencephalogram, EMG: electromyogram, EOG: electrooculogram (LOC/EOG1/E1, ROC/EOG2/E2: left, right EOG, respectively), LSTM: long short-term memory, MLP: multilayer perceptron, MSLT: multiple sleep latency test, PD: Parkinson's disease, PSG: polysomnography, RCNN: recurrent-convolutional neural network, RNN: recurrent neural network, VGG: visual geometry group. Acc.: Accuracy; Sens.: Sensitivity; Agr.: computer scoring and visual scoring agreement.

Authors	Dataset & Subjects	Channels	Classifier	Performance
Tsinalis et al. *2016* [79]	40 recordings (two PSG*20 healthy)[a]	EEG single-channel (Fpz-Cz)	CNNs + 2D stack of frequency-specific activity in time (end-to-end ANN)	Validation set *Overall Acc.* 71–76% *Per-stage Acc.* 80–84% *Min. Acc.* 60% N1 *Max. Acc.* 91% N3
Supratak et al. *2017* [27]	62 recordings (healthy) SS3[b] 40 recordings (two PSG*20 healthy)[a]	EEG single-channel (F4-EOG1 or Fpz-Cz or Pz-Oz)	Low-frequency information and high frequency information using CNNs + RNN (two bi-LSTM layers)	*Acc.* 86.2% *Kappa* 0.80 *Min. Sens.* 59.3% N1 *Max. Sens.* 90.7% N3 Validation set[a] *Acc.* 82.0% *Kappa* 0.76 *Min. Sens.* 50.1% N1 *Max. Sens.* 94.2% N3
Vilamala et al. *2017* [39]	40 recordings (two PSG*20 healthy)[a]	EEG single-channel (Fpz-Cz)	Time-frequency image + CNN (VGGNet as VGG-FE feature extractor and as VGG-FT fine-tuned network)	VGG-FE *Acc.* 84–88% *Min. Acc.* 68% N1 *Max. Acc.* 93% N3 VGG-FT *Acc.* 84–88% *Min. Acc.* 75% N1 *Max. Acc.* 94% N3
Biswal et al. *2018* [40]	10000 recordings[c] 5804 recordings[d]	EEG multi-channel (F3, F4, C3, C4, O1, O2) EEG multi-channel (C3, C4)	Spectrograms + RCNN (CNN + RNN)	Test set three *Acc.* 87.5% *Kappa* 0.80 *Min. Agr.* 58% N1 *Max. Agr.* 92% R Test set four *Acc.* 77.7% *Kappa* 0.73 *Min. Agr.* 48% N1 *Max. Agr.* 91% R
Chambon et al. *2018* [34]	61 recordings (healthy) SS3[b]	EEG multi-channel (F3, F4, C3, C4, O1, O2) EOG1, EOG2 three chin EMG	Multivariate network architecture: linear spatial filtering + CNN	Test set *Sens.* 52% *Min. Sens.* 58% N1 *Max. Sens.* 91% N3
Cui et al. *2018* [38]	116 recordings (healthy, sick, under treatment)[e]	EEG multi-channel (F3, F4, C3, C4, O1, O2) LOC, ROC X1, X2 and X3 EMG	CNN + fine-grained segment in multiscale entropy	Test set *Acc.* 92.2% *Min. Acc.* 86% N1 *Max. Acc.* 97% N3
Malafeev et al. *2018* [36]	54 recordings (three PSG*18 healthy)[f] 43 recordings (22 PSG and 21 MSLT narcolepsy and hypersomnia)[g]	EEG single-channel (Pz-Oz) one EMG two EOG	CNN (11 layers) + two bi-LSTM layers; Residual CNN (19 layers) + two bi-LSTM layers	Test set *Overall Kappa* 0.8 (except N1 with Kappa <0.5) *see paper for details*
Olesen et al. *2018* [75]	2310 recordings (healthy and patients)[h]	EEG multi-channel (central and occipital) EOG1, EOG2 chin EMG	Deep residual network model - 50 convolutional layers	Test set *Acc.* 84.1% *Kappa* 0.75 *Min. Sens.* 33.8% N3 *Max. Sens.* 93.8% W
Patanaik et al. *2018* [37]	1046 recordings DS1 (healthy adolescents)[i] 284 recordings DS2 (healthy young adults)[j] 210 recordings DS3 (sleep disorders)[k] 77 recordings DS4 (PD adults patients)[l]	EEG multi-channel (C3, C4) EOG (E1,E2)	Spectral Image + deep CNN + MLP stage classifier	Train on data[i,j] Test set[i,j] *Acc.* 89.8% *Kappa* 0.86 *Min. Acc.* 56.3% N1 *Max. Acc.* 94.3% W Validation set[k] *Acc.* 81.4% *Kappa* 0.74 *Min. Acc.* 40.9% N1 *Max. Acc.* 87.3% N2 Validation set[l] *Acc.* 72.1% *Kappa* 0.60 *Min. Acc.* 33.1% N1 *Max. Acc.* 80.1% N3

Table 1 (*continued*)

Authors	Dataset & Subjects	Channels	Classifier	Performance
Sors et al. *2018* [76]	5793 recordings (patients)[d]	EEG single-channel (C4)	14 layers CNN	Test set *Acc.* 87% *Kappa* 0.81 *Min. Sens.* 35% N1 *Max. Sens.* 91% W
Stephansen et al. *2018* [42]	3000 recordings (healthy and patients) from over 10 databases	EEG multi-channel (C3 or C4 and O1 or O2) LOC, ROC chin EMG	CNN + RNN	Test set on IS-RC *Acc.* 87% *see paper for details*
Zhang and Wu 2018 [33]	25 recordings (sleep-disordered breathing)[m] 16 recordings[n]	EEG single-channel	Complex-valued unsupervised CNN	Train on data[m] Validation set[m] *Acc.* 87% *Kappa* 0.81 *Min. Sens.* 80.2% N1 *Max. Sens.* 94.0% N2 Test set[n] *Acc.* 87.2%
Phan et al. *2019* [41]	200 recordings[b]	EEG single-channel (C4) EOG1, EOG2 two chin EMG	Time-frequency image + end-to-end hierarchical RNN for sequence-to-sequence sleep staging CNN	Test set *Acc.* 87.1% *Kappa* 0.81 *Min. Sens.* 59.7% N1 *Max. Sens.* 93.5% R
Yildirim et al. *2019* [32]	eight recordings (four healthy, four mild difficulty in falling asleep)[o] 61 recordings (healthy and mild difficulty in falling asleep)[a]	EEG single-channel (Fpz-Cz) single horizontal EOG channel	CNN	Test set 15 *Acc.* 91.22% Test set[a] *Acc.* 90.98%

Databases: Sleep-EDF*[Expanded].
[a] Montreal archive of sleep studies MASS* (SS1-SS5).
[b] Massachusetts General Hospital (MGH) Sleep Laboratory.
[c] Sleep Heart Health Study (SHHS)*.
[d] ISRUC-sleep*.
[e] University of Zurich.
[f] Psychiatry and Neurology in Warsaw.
[g] Wisconsin Sleep Cohort.
[h] CNL lab, Singapore.
[i] CSL lab, Singapore.
[j] SDU, Singapore GH.
[k] UC San Diego sleep lab.
[l] UCD database*.
[m] MIT-BIH database*.
[n] Sleep-EDF*.
[o] Accessible databases are identified by the asterisk (*) symbol.

Classifiers

Various classifiers have been used for automatic sleep scoring: CNN, deep neural networks (DNNs), RNN and even several combinations of them CNN + RNN or DNN + RNN. In most of the studies, CNN and RNN have been applied directly on raw PSGs data. Other approaches, that have shown a good performance, are based on the usage of precomputed spectrograms (spectral images representing the frequency content of the signals over time) combined with CNN and RNN [39,40,37,41].

During visual scoring, artifact removal is done using contextual information. Feature-based approaches are then more suitable for artifact identification and removal. In deep learning contextual information cannot be used, and the general aim is to produce a system ready to work with the data with a minimal manual preprocessing. Basic band pass filters (0.3–35 Hz) are usually applied, as recommended in the AASM. Few deep learning approaches consider artifact reduction. For example Cui et al. [38] use Butterworth filters, whilst Supratak et al. [27] apply a *weight decay* on the first layers of the CNNs in order to avoid overfitting to noises or artifacts in EEG data. Malafeev et al. [36] instead point out that deep learning algorithms can learn important information from epochs with artifacts, bringing as example the fact that wakefulness is almost always accompanied by movement artifacts and a movement is often followed by a transition into stage N1.

Deep learning can consider, thanks to the RNN, the temporal dynamics of the scoring procedure. Phan et al. [41] reached a significant performance improvement considering both the contextual input (sequence of multiple epochs) and the contextual output (sequence of multiple labels) during the learning process.

Performance

Performance is evaluated with a cross-fold validation process, using both k-fold and leave-one-out methodologies. In k-fold cross-validation, the PSG dataset is randomly partitioned into k equal sized recording groups. Out of the k groups a single group is retained as the validation data for testing the model; the remaining k-1 subgroups are used as training data. The cross-validation process is then repeated k times, with each of the k groups used exactly once as the validation data. The k results can then be averaged to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once. 10-fold cross-validation is commonly used. When k = n (the number of observations), the k-fold cross-validation is exactly the leave-one-out cross-validation.

Results are represented generally with the percentage agreement between the classifier and the gold standard, which is the visual scoring by a human expert. Accuracy and sensitivity are

usually given for each sleep stage, and overall values are also calculated. Accuracy and sensitivity resemble in general the same problematics as in the visual scoring procedure. N1 is the more difficult stage to be identified. The Cohen's kappa value is usually presented; it is generally thought to be a more robust measure as it takes into account the possibility of the agreement occurring by chance. The performance may be associated with training, validation and/or test set. The training set is the database partition used to develop the algorithm, so high performance is expected. Validation and test sets are both independent from the data with which the model has been built. As explained above, the validation set is used during the training phase, while the test set is used only to measure the final model performance. Big databases are usually divided into training, validation and test set, whilst smaller size databases are divided only into training and validation sets. As already said, the evaluation metrics are calculated considering the human visual scoring as the gold standard. Consequentially, a performance similar to the inter-scorer agreement (that is around 80%) can be considered an excellent result, whilst higher performance could be considered as overfitting on the dataset. In the cross-validation process the information of the training set and of the validation belongs to different recordings, but if they still belong to the same dataset, they cannot be considered totally independent. A dataset usually comes from the same sleep center and contains recordings from the same expert scorers. The high percentages of accuracy could be the result of an overfitting phenomenon (the models fit to the specific dataset). Data from different sleep labs and data from several cohorts ensure the reproducibility of the developed methods. Some authors have measured the performance of their model using test set coming from external database [40,42,37]; their results should be then considered more reliable.

Only one work has examined how the performance changes as a function of the epoch length, showing that the highest performance is obtained using a resolution of 30 s [42]. However they highlighted that the performance dropped only slightly with decreasing window sizes. It would then be possible to score sleep stages with high accuracy at lower time resolution (5 s), rendering the need for scoring per 30-second epochs obsolete.

It has also been demonstrated that performance usually increases with a multi-channel approach. In particular Chambon et al. [34] have shown that the accuracy improves employing up to six well distributed channels. They indicated that is worth adding more EEG sensors, but up to a certain point. Biswal et al. [40] showed that there is a small reduction in performance between six and two EEG channel approach, but still on par with the level of accuracy attained by experts. An alternative method to evaluate the performance of an algorithm could be to use the information contained in the hypnodensity graph presented by Stephansen et al. [42]. In this graph a probability distribution of each sleep stage is presented, conveying more information about the sleep trend.

Keeping in mind all the previous considerations, it appears quite clear the great difficulty of comparing different author works. All of them reach very good performance in terms of accuracy, compared with the inter-scorer agreement. In order to decide which classifier is better than the others, all the classifiers should be developed using the same channels, trained on the same dataset and validated with the same procedure.

Single-channel and EOG based automatic scoring

Generally, increasing independent information leads to an increase in the quality of the data analysis. Both feature-based [43], and raw signal approaches [34,38], have shown that multi-channel EEG together with EMG and EOG signal information leads to an increase in performance. The usage of an EEG multi-channel system improves the performance of the classification algorithms, and can also better gather sleep information. In fact, recent studies [44,45] have proven that sleep is not a global phenomenon, affecting the whole brain at the same time.

However, the improvement is usually quite small whilst adding more channel can be computationally expensive, and could compromise the efficiency of the algorithm without leading to a far better classification. Moreover many emerging home-based settings [46] require a reliable solution with few channels.

For these reasons, different groups have focused their attention on single-channel EEG or even on single-channel EOG analysis. The selection of the derivations is crucial, several debates can be found about the choice between Fpz-Cz and Pz-Oz electrodes, as the recent [47,48,35]. Several research groups have addressed the problem by extracting equally relevant information from frontal electrodes (Fp1-Fp2 EEG) [49–51].

Among the most recent EOG based works (EOG1-left or EOG2-right - single or double channels), Rahman et al. [52] method outperform in accuracy the previous ones [53–56]. The best performance, in a five-state classification algorithm, has been obtained using the SVM on Sleep-EDF database with an accuracy of about 93%. In this study even remarkable results have been obtained for the detection of the N1 stage (accuracy of about 65%) compared with the state-of-the-art of previous algorithms using single-channel EOG.

Semi-automated sleep scoring system – human intervention

All approaches that require a human intervention during the scoring can be considered semi-automated. The physician could be asked: to re-score the difficult part of the data (partial re-scoring); simply to pre-analyze the data prior to the automatic scoring; or to score the data following an assistive pre-analysis made by the computer. In Svetnik et al. work [57] different scoring procedures (manual, full re-scoring, partial re-scoring and fully automated) are compared, suggesting that protocols for partial re-scoring could be optimized to provide a high agreement with manual scoring at a reduced cost. In Koupparis et al. [58] a hypnospectrogram (whole-night time-frequency analysis) is presented to the human scorer to facilitate the staging decisions. They report a reduction of total time for sleep scoring of one third while still maintaining a substantial agreement between standard visual scoring and semi-automated analysis of the hypnospectrogram (Cohens kappa = 0.61). In Agarwal and Gotman work [59] an attempt is made to personalize the scoring to the single user. The user is asked to classify sample epochs, the method then learns from these samples to complete the classification, an overall agreement of around 80% is reached. All these approaches suggests the necessity of somehow inserting the human in the loop.

Latest products on the market

Table 2 lists the commercial products that, to the best of our knowledge, are currently present on the market and offer a system for automatic sleep scoring analysis. The search has been done on google with the following keywords: "*automatic sleep scoring products*", "*PSG and automatic sleep scoring products*", "*automated sleep scoring services and products*". Among all the outputs generated by the search, only the ones that pointed to a real company web page, and possibly with published research

Table 2
Scoring products actually available on the market. The first part of the table lists the products that offer the complete system, hardware and software, with both the PSG and the software tools for editing and executing the sleep scoring. The last part of the table lists the products that offer only the software.

Complete PSG systems and software		
Company name	Product name	Webpage link
Cidelec	Cidelec software	cidelec.net
Nihon Koden	Polysmith sleep systems	eu.nihonkohden.com
Natus Neuro	Embla Rem Logic	
SleepWorks		
Embla Sandman Elite	neuro.natus.com	
Somnomedics	Domino Diagnostic	somnomedics.de
Neurosoft	Neuron-Spectrum-PSG	neurosoft.com
Compumedics	Profusion Sleep	compumedics.com.au
Nox Medical	Noxturnal	noxmedical.com
CleveMed	Crystal PSG	clevemed.com
OSG	SleepRT	brainrt.com
Philips	Sleepware G3	philips.co.uk
Siesta Group	Somnolyzer	thesiestagroup.com
Sleep signals editing and automatic scoring software		
Company name	Product name	Webpage link
Michele	Michele Sleep Scoring	michelesleepscoring.com
WideMed	Morpheus	widemed.com
Neurobit Technologies	Z3Score	z3score.com

underneath the genesis of the product have been selected. They can be grouped in the ones that offer both the hardware needed for the PSG and the software for the following signal analysis and the products that provide only the software part. Nox Medical, Somnomedics and Siesta Group, are relatively young companies and are more focused on software for PSG analysis. The others listed in the first part of the table, instead, are historical producers of hardware and software for clinical applications, especially in the field of neurology. PSG hardware producers are also interested in providing adequate software for analyzing the collected data. Thus, it follows that to keep up with the times and market they had to give increasingly high-performance software to facilitate sleep physicians' tasks. Probably, also the continuous progress of research in the field of automatic sleep scoring has forced them to develop some proprietary solutions. Somnolyzer, Michele and Morpheus are the most widely referenced and tested in the literature. Several research studies present the performance of their algorithms and example of their application [60–64,57,65]. Finally, Z3Score appeared on the market, based on the work of Patanaik et al. [37]. Presently, almost all the market products for sleep scoring are developed with a feature-based approach, with the only exception of Z3Score that incorporates deep learning.

Automated PSG scoring and the clinical practice

In reviewing all the methodologies and approaches that have been challenged by easing and simplifying the scoring procedure, one question arises: why automated or at least semi-automated scoring is not already routinely adopted in all the sleep centers? In the following points the perspectives of ICT researchers and sleep scoring experts are summarized.

- ***Aversion to technology***: in the health care domain, new technologies are often perceived as a threat [66], especially if these new tools are going to substitute part of the work done by human beings and if they intervene somehow in the diagnostic process.
- ***Usability***: many tools that are on the market are not easy to use and have not a friendly user interface [67]; a user-centered design should be favored [68].
- ***Security and privacy issues***: some powerful scoring services require the up-loading of the sleep recordings data to the cloud, i.e., Z3Score [69], or externally to secure servers, i.e., Michele [70]. This action is often forbidden or discouraged by data protection policies of healthcare providers [71].
- ***Technical limitation***: automatic scoring works well on healthy subjects. Indeed, the majority of the machine learning approaches for improving sleep scoring have used a training set of healthy adult male subjects. Consequently applying these algorithms to patients with sleep disorders [16,36] or neurodegenerative disorders [37,72] often fails.
- ***Scoring rules***: the actual scoring rules leave space for subjective interpretation, leading to a high inter- and intra-scorer variability. Moreover the rules, based on 30-second epochs, tend to consider sleep stages as distinct entities, while sleep should be viewed as a gradual transition from a stage to the other [73]. Younes et al. [74] in a recent paper states that data interpretation performed by only one technologist should be considered unreliable.

Discussion

The machine learning algorithms and, in particular, the emerging deep learning approaches demonstrate a high level of accuracy and agreement – on average around 85% - between computer and visual scoring. Higher computational power has fostered the development of new approaches: learning directly from raw data can certainly have the advantage of revealing hidden information in the PSGs composite scoring procedure. In deep learning there is no need for *a priori* knowledge and for a concrete mathematical formulation, the learning process can be considered as a black-box. Many of the presented deep learning algorithms do not even consider an artifact removal process, letting the algorithm learn how to treat them. So possibly, given the complexity of sleep structure and of the scoring art, it may be the favorable approach. However, experimentally it does not provide always the best solution. Dataset characteristics, high bias and variance, play an important role in the accuracy reached.

In order to evaluate the application of deep learning in sleep scoring, performance is not enough. An overall accuracy comparable with the inter-scorer agreement is of course the minimal requirement. Extremely high accuracy should be evaluated carefully, high values are related often to data overfitting. Certainly, dataset characteristics play an important role. The dataset has not to be simply big. High heterogeneity in the dataset is needed, in order to resemble the high heterogeneity in the human sleep and in the human scoring. Training an algorithm on a dataset with few scorers from the same lab and with PSGs belonging only to healthy young males may not reproduce the performance on a test set coming from a different database. In this perspective, works using a high number of recordings [40,75,37,76,42], assume a greater value given the heterogeneity of the subjects and given the deep learning prerequisites.

Despite the advantage of scoring data from many EEG channels, very good performance has been achieved with a single EEG channel approach, or even with a single EOG channel. Single-channel is indeed a very promising approach, but it should be limited to home-based solution, for early diagnosis and continuous monitoring. Considering the complexity of sleep itself and of sleep disorders the information of many channels is still essential for a rigorous scoring. Deep learning is computationally expensive. Increasing the number of channels to be considered for scoring and

sleep characterization compromises the efficiency of the algorithm. Future deep neural networks should try to minimize computational costs, possibly emulating brain strategies as indicated by Schmidhuber [77]. Also, existing semi-automatic approaches, where the human intervention during the scoring is required, seem to be a promising way forward.

Nevertheless, all these approaches fail to be introduced in the daily routine. It seems that the real needs of the physicians are still not completely understood. The visual scoring rules have limitations that are in contrast with the actual computational possibilities. For example, they are still tight to fixed length time epochs and to visual identification of complex patterns. Sleep is a continuous process, whilst staging is a stepwise process. Several continuous digital markers of sleep depth have been proposed to improve the rules, mainly related to the power like for example the odds ratio product [78], and possibly will be considered in updated version of the AASM standards. It is really challenging to develop an automatic scoring system that follows the standards while the standards are still tight to technical limitations that have to be overcome.

Deep learning has the enormous power of extracting hidden information from the PSG recordings and from the scoring procedure. RNN can consider contextual information (sequence of multiple epochs and of multiple stages) as the human scorer does. The classifier can even extract information from the artifacts, without the need for their removal. These algorithms resemble possibly in a better way the human reasoning.

Conclusion

In this review the latest application of deep learning in automatic sleep scoring have been thoroughly examined, comparing them with other approaches. Deep learning methodologies present many advantages. Algorithms can be applied directly on raw data, with minimal artifact removal. Hidden information, ignored by features approaches, can be revealed, as well as information related both to the PSG signal and to the human scoring procedure itself. While doing the scoring, the physician may consider, unconsciously, information like patient demographics, medication, health status, suspected disorder, already seen patterns and fuses this information within the limits of the AASM. Unconscious information is difficult to be characterized and to be translated in a feature-based approach, whilst deep learning can learn also human perception. However, this black-box behavior is also the main drawback, as it could limit the end-user acceptance. In a feature-based approach the user has clear information about the sleep characteristic considered for scoring, while in deep learning this is not a straightforward. The original goal of scoring rules to simplify the procedure and to harmonize sleep laboratories analysis. However, adopting fixed-time stages leads to an artificial scoring that may not be able to evaluate the epoch transitions correctly. Strict sleep stages of a hypnogram are a convention between researchers rather than a physiological truth. The subjectivity in the interpretation of the classification scheme by different scorers is unavoidable, as stated already in [59]. Nevertheless, this highly debated convention allows the evaluation of human sleep and the identification of sleep disorders.

New research in deep learning should focus on artifact removal, without losing significant pattern, and on how to increase channel information without compromising the stability of the algorithms. Special attention during performance evaluation has to be reserved to the dataset that has been used for development and for testing. Alternative methods for evaluation, like the proposed hypnodensity graph, should be considered. A highly heterogeneous dataset is recommended. Ideally, in order to compare different classifiers, they should be trained and tested on the same datasets. Summarizing, deep learning is most probably the more promising approach for automated scoring. A multidisciplinary approach that takes into account real end-user needs and the engineering perspective has the potential to develop a reliable and well accepted automatic scoring system.

Practice points

1. Machine learning and deep learning algorithms have been exhaustively used for ASSC; high levels of accuracy comparable with the inter-scorer agreement rate can be reached.
2. Sequential deep learning algorithms applied on raw data have shown to be promising approaches, since sequences over time are stored and are taken into account during the manual sleep staging procedure.
3. Algorithm goodness should be evaluated considering not only the performance but also dataset characteristics, validation and testing methods.

Research agenda

1. A better-defined artifact countermeasure, a higher heterogeneity in the dataset and a stable capability for multi-channel information have the potential to lead to further improvement in scoring by deep learning.
2. Single-channel scoring methods are promising, their use in a pre-diagnostic home-based assessment and for continuous monitoring should be properly investigated and validated.

Conflicts of interest

There is no conflict of interest.

Acknowledgements

We have conducted several short interviews to gather feedback on sleep scoring problematics and possible improvements. We would like to thank Dr. med. Manuel Sastry (Academic Sleep Clinic, CIRO, Horn, The Netherlands), Dr. med. Pierre-Alois Beitinger (Sleep Laboratory Max-Planck-Institut für Psychiatrie, Munich, Germany) and especially Dr. med. Mauro Manconi (Sleep and Epilepsy Center, Neurocenter of Southern Switzerland, Civic Hospital (EOC), Lugano, Switzerland). Prof. P. Favaro was supported by the IRC Decoding Sleep: From Neurons to Health and Mind, from the University of Bern, Switzerland.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.smrv.2019.07.007.

References

[1] Ohayon MM. Epidemiological overview of sleep disorders in the general population. Sleep Med Res (SMR) 2011;2(1):1–9.
[2] Penzel T, Conradt R. Computer based sleep recording and analysis. Sleep Med Rev 2000;4(2):131–48.
[3] Rosenberg RS, Van Hout S. The American academy of sleep medicine inter-scorer reliability program: sleep stage scoring. J Clin Sleep Med 2013;9(01):81–7.
[4] Younes M, Raneri J, Hanly P. Staging sleep in polysomnograms: analysis of inter-scorer variability. J Clin Sleep Med 2016;12(06):885–94.
[5] Muto V, Berthomier C, Schmidt C, Vandewalle G, Jaspar M, Devillers J, et al. 0315 Inter-and intra-expert variability in sleep scoring: comparison between visual and automatic analysis. Sleep 2018;41(suppl_1):A121.
[6] Rechtschaffen A, Kales A. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. US Government Printing Office, US Public Health Service; 1968.
[7] Iber C, Iber C. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications, vol. 1. Westchester, IL: American Academy of Sleep Medicine; 2007.
[8] Danker-hopfe H, Anderer P, Zeitlhofer J, Boeck M, Dorn H, Gruber G, et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. J Sleep Res 2009;18(1):74–84.
[9] Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med 2009;6(7):e1000100.
*[10] Motamedi Fakhr S, Moshrefi-Torbati M, Hill M, Hill CM, White PR. Signal processing techniques applied to human sleep EEG signals-A review. Biomed Signal Process Control 2014;10:21–33.
[11] Şen B, Peker M, Çavuşoğlu A, Çelebi FV. A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms. J Med Syst 2014;38(3):18.
*[12] Aboalayon K, Faezipour M, Almuhammadi W, Moslehpour S. Sleep stage classification using EEG signal analysis: a comprehensive survey and new investigation. Entropy 2016;18(9):272.
[13] Lan T. Feature extraction feature selection and dimensionality reduction techniques for brain computer interface. 2011. Available from: https://scholararchive.ohsu.edu/concern/etds/np193924b?locale=pt-BR.
*[14] Ronzhina M, Janoušek O, Kolářová J, Nováková M, Honzík P, Provazník I. Sleep scoring using artificial neural networks. Sleep Med Rev 2012;16(3):251–63.
[15] Radha M, Garcia-Molina G, Poel M, Tononi G. Comparison of feature and classifier algorithms for online automatic sleep staging based on a single EEG signal. In: 2014 36th Annual International Conference of the IEEE engineering in medicine and biology society. IEEE; 2014. p. 1876–80.
*[16] Boostani R, Karimzadeh F, Nami M. A comparative review on sleep stage classification methods in patients and healthy individuals. Comput Methods Progr Biomed 2017;140:77–91.
[17] Güneş S, Polat K, Yosunkaya Ş. Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting. Expert Syst Appl 2010;37(12):7922–8.
[18] Acharya UR, Chua ECP, Chua KC, Min LC, Tamura T. Analysis and automatic identification of sleep stages using higher order spectra. Int J Neural Syst 2010;20(06):509–21.
[19] Fraiwan L, Lweesy K, Khasawneh N, Fraiwan M, Wenz H, Dickhaus H. Classification of sleep stages using multi-wavelet time frequency entropy and LDA. Methods Inf Med 2010;49(03):230–7.
[20] Fraiwan L, Lweesy K, Khasawneh N, Wenz H, Dickhaus H. Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier. Comput Methods Progr Biomed 2012;108(1):10–9.
[21] Liang SF, Kuo CE, Hu YH, Pan YH, Wang YH. Automatic stage scoring of single-channel sleep EEG by using multiscale entropy and autoregressive models. IEEE Trans Instrum Meas 2012;61(6):1649–57.
[22] Biswal S, Kulas J, Sun H, Goparaju B, Westover MB, Bianchi MT, et al. SLEEPNET: automated sleep staging system via deep learning. CoRR 2017. abs/1707.08262. Available from: http://arxiv.org/abs/1707.08262.
[23] Dong H, Supratak A, Pan W, Wu C, Matthews PM, Guo Y. Mixed neural network approach for temporal sleep stage classification. IEEE Trans Neural Syst Rehabil Eng 2018;26(2):324–33.
[24] Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L. Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer vision and pattern recognition; 2014. p. 1725–32.
[25] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems; 2012. p. 1097–105.
[26] Hannun AY, Case C, Casper J, Catanzaro B, Diamos G, Elsen E, et al. Deep Speech: scaling up end-to-end speech recognition. CoRR 2014. abs/1412.5567, http://arxiv.org/abs/1412.5567.
[27] Supratak A, Dong H, Wu C, Guo Y. DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG. IEEE Trans Neural Syst Rehabil Eng 2017;25(11):1998–2008.
[28] PhysioNet, The Sleep-EDF Database. Available from: https://www.physionet.org/physiobank/database/sleep-edf/ [accessed on 19 November 2018].
[29] PhysioNet, The Sleep-EDF (Expanded) Database. Available from: https://www.physionet.org/physiobank/database/sleep-edfx/ [accessed on 19 November 2018].
[30] O'Reilly C, Gosselin N, Carrier J, Nielsen T. Montreal Archive of Sleep Studies: an open-access resource for instrument benchmarking and exploratory research. J Sleep Res 2014;23(6):628–35. Available from: https://massdb.herokuapp.com/en/.
[31] Quan SF, Howard BV, Iber C, Kiley JP, Nieto FJ, O'Connor GT, et al. The sleep heart health study: design, rationale, and methods. Sleep 1997;20(12):1077–85. Available from: https://sleepdata.org/datasets/shhs.
[32] Yildirim O, Baloglu UB, Acharya UR. A deep learning model for automated sleep stages classification using PSG signals. Int J Environ Res Public Health 2019;16(4):599.
[33] Zhang J, Wu Y. Complex-valued unsupervised convolutional neural networks for sleep stage classification. Comput Methods Progr Biomed 2018;164:181–91.
[34] Chambon S, Galtier MN, Arnal PJ, Wainrib G, Gramfort A. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. IEEE Trans Neural Syst Rehabil Eng 2018;26(4):758–69.
[35] Tsinalis O, Matthews PM, Guo Y. Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders. Ann Biomed Eng 2016;44(5):1587–97.
[36] Malafeev A, Laptev D, Bauer S, Omlin X, Wierzbicka A, Wichniak A, et al. Automatic human sleep stage scoring using deep neural networks. Front Neurosci 2018;12:781.
*[37] Patanaik A, Ong JL, Gooley JJ, Ancoli-Israel S, Chee MW. An end-to-end framework for real-time automatic sleep stage classification. Sleep 2018;41(5):zsy041.
[38] Cui Z, Zheng X, Shao X, Cui L. Automatic sleep stage classification based on convolutional neural network and fine-grained segments. Complexity 2018;2018.
[39] Vilamala A, Madsen KH, Hansen LK. Deep convolutional neural networks for interpretable analysis of EEG sleep stage scoring. In: 2017 IEEE 27th International workshop on machine learning for signal processing (MLSP). IEEE; 2017. p. 1–6.
[40] Biswal S, Sun J, Sun H, Westover MB, Goparaju B, Bianchi MT. Expert-level sleep scoring with deep neural networks. J Am Med Inf Assoc 2018;25(12):1643–50.
[41] Phan H, Andreotti F, Cooray N, Chn OY, De Vos M. SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. IEEE Trans Neural Syst Rehabil Eng 2019:1.
*[42] Stephansen JB, Olesen AN, Olsen M, Ambati A, Leary EB, Moore HE, et al. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. Nat Commun 2018;9(1):5229.
[43] Melia U, Guaita M, Vallverdú M, Embid C, Vilaseca I, Salamero M, et al. Mutual information measures applied to EEG signals for sleepiness characterization. Med Eng Phys 2015;37(3):297–308.
[44] Siclari F, Bassetti C, Tononi G. Conscious experience in sleep and wakefulness. Swiss Arch Neurol Psychiatr 2012;163:273–8.
[45] Nir Y, Staba RJ, Andrillon T, Vyazovskiy VV, Cirelli C, Fried I, et al. Regional slow waves and spindles in human sleep. Neuron 2011;70(1):153–69.
[46] Van De Water AT, Holmes A, Hurley DA. Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography–a systematic review. J Sleep Res 2011;20(1pt2):183–200.
[47] Hsu YL, Yang YT, Wang JS, Hsu CY. Automatic sleep stage recurrent neural classifier using energy features of EEG signals. Neurocomputing 2013;104:105–14.
[48] Zhu G, Li Y, Wen PP. Analysis and classification of sleep stages based on difference visibility graphs from a single-channel EEG signal. IEEE J Biomed Health Inf 2014;18(6):1813–21.
[49] Stepnowsky C, Levendowski D, Popovic D, Ayappa I, Rapoport DM. Scoring accuracy of automated sleep staging from a bipolar electroocular recording compared to manual scoring by multiple raters. Sleep Med 2013;14(11):1199–207.
[50] Huang CS, Lin CL, Ko LW, Liu SY, Su TP, Lin CT. Knowledge-based identification of sleep stages based on two forehead electroencephalogram channels. Front Neurosci 2014;8:263.
[51] Popovic D, Khoo M, Westbrook P. Automatic scoring of sleep stages and cortical arousals using two electrodes on the forehead: validation in healthy adults. J Sleep Res 2014;23(2):211–21.
*[52] Rahman MM, Bhuiyan MIH, Hassan AR. Sleep stage classification using single-channel EOG. Comput Biol Med 2018;102:211–20.
[53] Olesen AN, Christensen JA, Sorensen HB, Jennum PJ. A noise-assisted data analysis method for automatic EOG-based sleep stage classification using ensemble learning. In: 2016 38th Annual International Conference of the IEEE engineering in medicine and biology society (EMBC). IEEE; 2016. p. 3769–72.
[54] Xia B, Li Q, Jia J, Wang J, Chaudhary U, Ramos-Murguialday A, et al. Electrooculogram based sleep stage classification using deep belief network. In:

* The most important references are denoted by an asterisk.

2015 International joint Conference on neural networks (IJCNN). IEEE; 2015. p. 1–5.
[55] Liang SF, Kuo CE, Lee YC, Lin WC, Liu YC, Chen PY, et al. Development of an EOG-based automatic sleep-monitoring eye mask. IEEE Trans Instrum Meas 2015;64(11):2977–85.
[56] Virkkala J, Hasan J, Värri A, Himanen SL, Müller K. Automatic sleep stage classification using two-channel electro-oculography. J Neurosci Methods 2007;166(1):109–15.
[57] Svetnik V, Ma J, Soper KA, Doran S, Renger JJ, Deacon S, et al. Evaluation of automated and semi-automated scoring of polysomnographic recordings from a clinical trial using zolpidem in the treatment of insomnia. Sleep 2007;30(11):1562–74.
[58] Koupparis AM, Kokkinos V, Kostopoulos GK. Semi-automatic sleep EEG scoring based on the hypnospectrogram. J Neurosci Methods 2014;221:189–95.
[59] Agarwal R, Gotman J. Computer-assisted sleep staging. IEEE (Inst Electr Electron Eng) Trans Biomed Eng 2001;48(12):1412–23.
*[60] Anderer P, Gruber G, Parapatics S, Woertz M, Miazhynskaia T, Klösch G, et al. An E-health solution for automatic sleep classification according to Rechtschaffen and Kales: validation study of the Somnolyzer 24×7 utilizing the Siesta database. Neuropsychobiology 2005;51(3):115–33.
*[61] Anderer P, Moreau A, Woertz M, Ross M, Gruber G, Parapatics S, et al. Computer-assisted sleep classification according to the standard of the American Academy of Sleep Medicine: validation study of the AASM version of the Somnolyzer 24×7. Neuropsychobiology 2010;62(4):250–64.
[62] Punjabi N M, Shifa N, Doffner G, Patil S, Pien G, Aurora R. Computer-assisted automated scoring of polysomnograms using the somnolyzer system. Sleep 2015;38.
[63] Younes M. The case for using digital EEG analysis in clinical sleep medicine. Sleep Sci Pract 2017;1(2).
[64] Malhotra A, Younes M, Hanlon A, Staley B, Pien GW, Pack AI, et al. Performance of an automated polysomnography scoring system versus computer-assisted manual scoring. Sleep 2013 04;36(4):573–82.
[65] Pittman S, MacDonald M, Fogel R, Malhotra A, Todros K, Levy B, et al. Assessment of automated scoring of polysomnographic recordings in a population with suspected sleep-disordered breathing. Sleep 2004 12;27:1394–403.
[66] Fichman RG, Kohli R, Krishnan R. Editorial overview-the role of information systems in healthcare: current research and future trends. Inf Syst Res 2011;22(3):419–28.
[67] Marcilly R, Peute L, Beuscart-Zephir MC. From usability engineering to evidence-based usability in health IT. Stud Health Technol Inform 2016;222:126–38.
[68] Kushniruk A, Nøhr C. Participatory design, user involvement and health IT evaluation. Stud Health Technol Inform 2016;222:139–51.
[69] Tay J, Toh S, Leow L, Senin S. Assessing competency of Z3Score automated sleep stage scoring system with manual sleep stage scoring by multiple scorers. Sleep Med 2017;40:e326.
[70] Younes M, Thompson W, Leslie C, Egan T, Giannouli E. Utility of technologist editing of polysomnography scoring performed by a validated automatic system. Ann Am Thorac Soc 2015;12(8):1206–18.
[71] Ali O, Shrestha A, Soar J, Wamba SF. Cloud computing-enabled healthcare opportunities, issues, and applications: a systematic review. Int J Inf Manag 2018;43:146–58.
[72] Jensen PS, Sorensen HB, Leonthin HL, Jennum P. Automatic sleep scoring in normals and in individuals with neurodegenerative disorders according to new international sleep scoring criteria. J Clin Neurophysiol 2010;27(4):296–302.
[73] Malhotra RK, Avidan AY. Introduction to sleep stage scoring. Atlas Sleep Med 2013:77.
*[74] Younes M, Kuna ST, Pack AI, Walsh JK, Kushida CA, Staley B, et al. Reliability of the American academy of sleep medicine rules for assessing sleep depth in clinical practice. J Clin Sleep Med 2018;14(02):205–13.
[75] Olesen AN, Jennum P, Peppard P, Mignot E, Sorensen HB. Deep residual networks for automatic sleep stage classification of raw polysomnographic waveforms. In: 2018 40th Annual International Conference of the IEEE engineering in medicine and biology society (EMBC). IEEE; 2018. p. 1–4.
[76] Sors A, Bonnet S, Mirek S, Vercueil L, Payen JF. A convolutional neural network for sleep stage scoring from raw single-channel EEG. Biomed Signal Process Control 2018;42:107–14.
[77] Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw 2015;61:85–117.
[78] Younes M, Ostrowski M, Soiferman M, Younes H, Younes M, Raneri J, et al. Odds ratio product of sleep EEG as a continuous measure of sleep state. Sleep 2015;38(4):641–54.
[79] Tsinalis O, Matthews PM, Guo Y, Zafeiriou S. Automatic sleep stage scoring with single-channel EEG using convolutional neural networks. arXiv e-prints; 2016 Oct. p. arXiv:1610.01683.