Informatics in Medicine Unlocked 20 (2020) 100370

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: http://www.elsevier.com/locate/imu

Investigating the use of uni-directional and bi-directional long short-term memory models for automatic sleep stage scoring

Luay Fraiwan [a,b,*], Mohanad Alkhodari [a]

[a] Department of Electrical and Computer Engineering, Abu Dhabi University, Abu Dhabi, United Arab Emirates
[b] Department of Biomedical Engineering, Jordan University of Science and Technology, Irbid, Jordan

ARTICLE INFO

Keywords:
Sleep stage scoring
Deep learning
Recurrent neural network
Long short-term memory
Training
Classification

ABSTRACT

In this paper, a study is conducted to investigate the use of a Long Short-Term Memory (LSTM) learning system in automatic sleep stage scoring. The developed algorithm will automatically learn to classify sleep stages from any acquired sleep signals data-set. This allows to resolve the difficulties that are facing experts in manual sleep stage scoring. A total of 39 Polysomnogram (PSG) recordings acquired from the online PhysioNet Sleep-EDF database are used in this study. The PSG recordings are chosen to be from the EEG Fpz-Cz signals only. The database comes with annotation files that include expert manual stage scoring based on the Rechtschaffen & Kales (R&K) scoring manual. The obtained signals go initially through a pre-processing procedure where sleep stages signals are extracted, normalized, and filtered. The resulting sleep signals are trained using a k-fold cross-validation scheme of 10-folds. Prior to the training and classification process, the LSTM network architecture is built using Uni- and Bi-directional structures to utilize both the forward and backward chains of data sequences. At the end, the developed algorithm performance is evaluated and a complete performance summary table is provided relative to other State-of-the-Art deep learning studies. The performance of this study is evaluated initially without the merging of S3 and S4 sleep stages following the R&K manual, which is considered challenging due to the minor differences between the signals. Then, the performance is evaluated following the recent American Academy of Sleep Medicine (AASM) scoring manual with the merging of the two stages as N3. The developed algorithm achieved higher results using the Bi-directional LSTM. In addition, it achieved the highest accuracy among all other studies in the field with 97.28%. Furthermore, Cohen's kappa and F1-score were more than 72% on average between all sleep stages. According to the confusion matrix, the algorithm successfully classified sleep signals with an overall True Positives percentage of 91.92%. The performance of the algorithm improved following the AASM manual, where the Cohen's kappa value increased from 72.55% to 77.73%. The developed algorithm showed potential in automatic sleep stage classification. Future works include further enhancements on the LSTM algorithm to achieve higher levels of performance.

1. Introduction

Sleep is a naturally restorative state for the health of the human mind and body. It is mainly characterized by a suppression of all voluntary muscles with an altered consciousness, where humans go under a mental and physical recurring process [1]. However, a major problem occurs when this natural state is interrupted by various sleep disorders such as sleep apnea, insomnia, and restless legs syndrome [2–4]. According to the World Health Organization (WHO), nearly 1 Billion people suffers from Obstructive Sleep Apnea (OSA) worldwide [5]. In addition, the American Academy of Sleep Medicine (AASM) estimated that more than 30% of adults at the US have symptoms of insomnia [6]. The gold standards to diagnose human sleep state is the Polysomnogram (PSG) [7–9]. This laboratory tool monitors the sleep overnight based on brain activity (Electroencephalogram, EEG), heart rhythm (Electrocardiogram, ECG), muscle activation (Electromyogram, EMG), and eye movement (Electrooculogram, EOG) biosignals [10–12]. It assists clinicians to manually determine sleep stages and scores of patients by labeling each segment of 20–30 s duration, called an epoch, with its corresponding sleep stage [9,12].

Sleep stage scoring is a technique that is commonly used to identify sleep stages and help in sleep disorders treatment [12]. It is normally

* Corresponding author. Department of Electrical and Computer Engineering, Abu Dhabi University, Abu Dhabi, United Arab Emirates.
E-mail addresses: fraiwan@just.edu.jo (L. Fraiwan), 1045804@students.adu.ac.ae (M. Alkhodari).

https://doi.org/10.1016/j.imu.2020.100370
Received 8 April 2020; Received in revised form 8 June 2020; Accepted 8 June 2020
Available online 20 June 2020

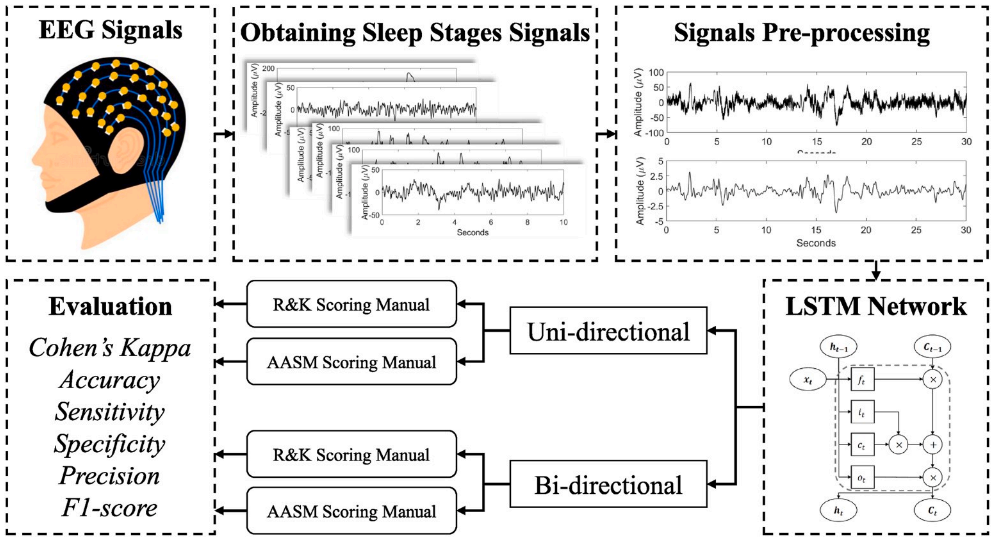

Fig. 1. The complete procedure followed in the study.

done according to two manuals: the old Rechtschaffen & Kales (R&K) and the recently introduced AASM. According to R&K scoring manual [13], sleep is divided into six stages: an Awake (W) stage followed by four stages referred to as the Non-Rapid Eye Movement (NREM) stage and ends by the REM stage. The NREM stage includes the transitional sleep (S1), light sleep (S2), deep sleep (S3), and slow wave sleep (S4) stages [1,14]. In 2007, AASM produced an updated scoring manual [15]. The stage W (same as R&K) stage is represented by the alpha waves of frequency 8–12 Hz. Stage N1 (S1 in R&K) is associated with slower theta waves of frequency 4–7 Hz. Stage N2 (S2 in R&K) is usually linked with an increased amplitude of the signals, sleep spindles of 12–14 Hz frequency bursts of activity, and K-complex fast waves. In stage N3 (S3 and S4 in R&K), the delta wave of frequencies from 0 to 3 Hz dominates, while the sleep spindles and K-complex continue to appear but with a decreased amplitude. The final stage, R (REM in R&K), is characterized by having dreams and a combinations of alpha, beta, and theta waves [14,16,17]. Based on these classifications, clinicians up to date manually diagnose acquired sleep PSG recordings. This manual sleep staging and scoring process is considered convoluted and time consuming, as there may be multiple observations and conclusions from different clinicians due to signals artifacts [1,18]. In summary, the major differences between the two manuals in scoring are: the difference in terminology and the combination of S3 and S4 into one stage N3. Further differences exists concerning the derivation of electrodes, the definition of sleep-wake transition, slow waves, k-complexes, etc.

Recently, there has been a wide interest to provide an accurate automated sleep staging process to assist in sleep disorders diagnosis. Most automated sleep stages classifications rely on either feature extraction algorithms or deep learning networks [3,12,19]. In feature extraction, features that could be extracted from time- or frequency-domain signals are used to train a classifier to automatically predict the stages. In this area, many research works focused on feature extraction using fuzzy classification [20,21], Support Vector Machine (SVM) [22–26], wavelet transform [27–29], decision trees [24,30,31], random forest [27,32,33], and nearest neighbours algorithms [34]. These works showed a strong potential of using hand-crafted features for sleep stage scoring. However, due to the variable nature of the PSG recordings, these methods may lack the needed accuracy and require continuous tuning. In deep learning networks, this issue is solved by training the system to learn directly from the raw recordings data [3]. Research works through deep learning are focusing mainly on the use of Convolutional Neural Networks (CNNs) [35–37] and Recurrent Neural Networks (RNNs) [38–43] in sleep stage scoring.

In Artificial Intelligence (AI), RNN is a neural network that is characterized by having a chain-like structure. The connections between nodes form dynamic loops to allow for capturing the temporal sequence of data [44,45]. Even though the use of RNN showed efficient outcomes, it is considered difficult to train the network due to the exploding gradient problems of the back-propagating process [45]. To overcome this issue, Long Short-Term Memory (LSTM) networks are proposed to utilize the long-term dependencies of data. LSTM networks include a feedback connection that allows gradients to flow unchanged. It has been used for various applications including image captioning, speech recognition, and language modeling [46,47]. By incorporating LSTM within the framework of RNN for biomedical applications, sleep stage scoring becomes achievable.

1.1. Our contribution

In this paper, a study is conducted on the use of LSTM to automatically classify sleep stages from acquired PSG recordings. A single raw EEG signal (Fpz-Cz) is extracted from PSG recordings along with the corresponding sleep stage. The EEG signal is divided into epochs of 30 s and annotated with its corresponding sleep stage prior to pre-processing steps of normalization and digital filtering. The stages are selected initially based on the R&K scoring manual, then they are changed following the recent AASM scoring manual to compare the performance. The pre-processed EEG signal is used as a feature vector to be fed to an LSTM classification network following both the Uni- and Bi-directional structures. Prior to the classification and training process, the resulting sleep signals are divided into training and testing following a k-fold cross-validation scheme of 10-folds. In addition, the LSTM network architecture is built with a pre-defined training parameters to ensure an improved performance. At the end, the overall performance of the developed algorithm is evaluated using multiple metrics such as the accuracy, sensitivity, specificity, precision, F1-score, and Cohen's Kappa (κ). Furthermore, a summary table is provided to report the performance of this study relative to other State-of-the-Art studies that use deep learning for automated sleep stage scoring.

The main contribution of this work is to investigate the use of Long Short-Term Memory models in classifying sleep stages corresponding to the two current scoring manuals. In addition, reporting the efficiency of using Uni-directional and Bi-directional LSTM for automated classification based on LSTM-based extracted features rather than manual feature engineering. Despite achieving high levels of accuracy, manual feature engineering may be sensitive to errors. Extracting features manually could be very sensitive to noise, strength of the acquired signals, and other manual adjustments done prior to the sleep identification process. These manual adjustments are omitted by using neural networks, which usually provide better performance but with the price of a higher computational demand as well as more training data to be included in the learning model [48]. Therefore, LSTM layers learn features and utilize them to classify signals to their corresponding stages. The models learn each signal features on its own based on its conditions without any manual features interference. Furthermore. the S3 and S4 sleep stages are separated and the performance of the algorithm is tested unlike most studies where they merge both stages prior to the classification process. Furthermore, a summary table provided a short review of the current research works in the field. The overall procedure followed in the study is illustrated in Fig. 1.

1.2. Paper organizations

The paper is organized as follows. In Section 2, a brief explanation of the architecture of LSTM is provided along with the corresponding mathematical formulation in both the forward and backward chains. Section 3, the materials and methods used in this study are outlined including the selected subjects database, pre-processing steps, LSTM classification and training process, and performance evaluation metrics explanation. For Section 4, the overall performance of the algorithm is evaluated and discussed, In addition, a complete summary table of this study's performance relative to other State-of-the-Art studies is provided. The paper is concluded with future works in Section 5.

2. LSTM architecture

The architecture of the LSTM neural network is composed of a cell surrounded by multiple units that work as memory blocks. These blocks store the temporal state of the network continuously. The cell works by implementing these blocks, named as gates, to manage the information flow within the network [49,50]. These gates are the input (i), output (o), and forget (f) gates. The input gate controls the activations of the input flow, the output gate controls the activation of the output flow, and the forget gate resets the memory to calculate the data proportion needed to be kept [49,51]. To be able to provide a short-term memory storage, the Constant Error Carousel (CEC) recirculates the activations and the error signals each time an input is entering the network. Furthermore, peephole connections are used as the feedback sources between the cell and the gates [52].

In terms of mathematical equations, conventional RNN output vector y and hidden vector h at time instant t are usually updated iteratively based on the input sequence $x = (x_1, x_2, \ldots, x_T)$ as follows,

$$h_t = \mathscr{H}_h(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{1}$$

$$y_t = \mathscr{H}_y(W_{hy}h_t + b_y) \tag{2}$$

where $\mathscr{H}_h$ is the hidden layer function, W is the weight matrix between vectors, b is the bias vector [53,54], and $\mathscr{H}_y$ is the output layer function.

However, as previously discussed, upon training the system with this structure, the exploding gradient problem appears when back-propagation takes place. Therefore, the LSTM architecture allows for resolving this issue. The standard structure of the LSTM network in the forward chain is described by the following equations to better implement the hidden layer $\mathscr{H}$ function,

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{3}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{4}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{5}$$

$$c_t = tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{6}$$

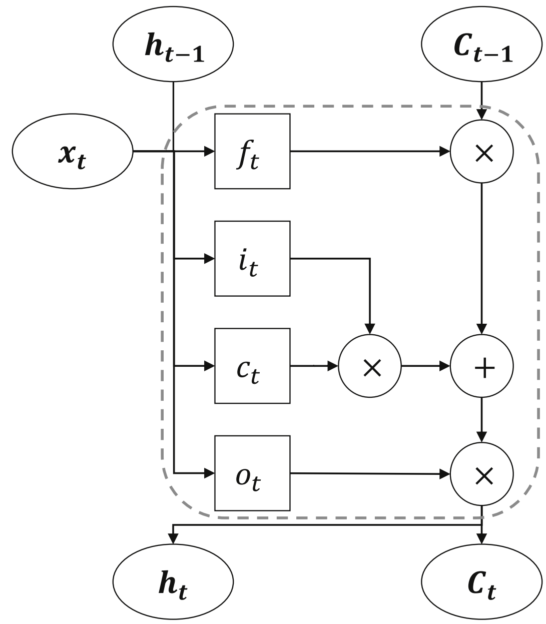

Fig. 2. The Standard Long Short-Term Memory (LSTM) architecture in the forward direction.

$$C_t = f_tC_{t-1} + i_tc_t \tag{7}$$

$$h_t = o_t tanh(c_t) \tag{8}$$

where W_{x*} are the input-to-gate weights, W_{h*} are the hidden-to-hidden weights, W_{c*} are the peephole weights, c is the cell input, C is the cell output, and h is the hidden layer output. The function used is the sigmoid, $\sigma()$, representing a hard gating function bounded by (0,1). The complete architecture is shown in Fig. 2. The dashed line represents the LSTM cell and gates.

On the backward chain, the hidden layer h_{t+1} is utilized in the equations [49,51]. To process sequence data in both the forward and backward chains, the Bi-directional LSTM (BDLSTM) networks are introduced [51,55]. The final output, y_t, of a BDLSTM network is represented as,

$$y_t = W_{\overrightarrow{h}y}\overrightarrow{h^N} + W_{\overleftarrow{h}y}\overleftarrow{h^N} + b_y \tag{9}$$

where $\overrightarrow{h^N}$ and $\overleftarrow{h^N}$ are the hidden layers output in the forward and backward directions, respectively, for all N levels of stack. Both hidden layers outputs are calculated iteratively using positive and reverse sequences from time $t-n$ to $t-1$, respectively, where n is the historical time frames or steps. Both outputs are then concatenated using a sigma (σ) function and fed to the main output vector. The forward chain equations use hidden layers of h_{t-1}, whereas the backward chain equations use hidden layers of h_{t+1}.

3. Material and methods

3.1. Subjects

The EEG data-set is obtained from the PhysioNet Sleep-EDF Database [56] that has been widely used in literature [57–60]. The database includes a collection of two PSG recordings data-sets obtained for two separate studies. The first data-set, named as *Sleep Cassette (SC)*, contains 39 PSG recordings acquired for a study on the effects of age on sleep patterns between 1987 and 1991. The subjects of the study are healthy 20 Caucasian patients (10 Male and 10 Female) of 25–35 years old not taking any type of medications. From each patient, two PSG

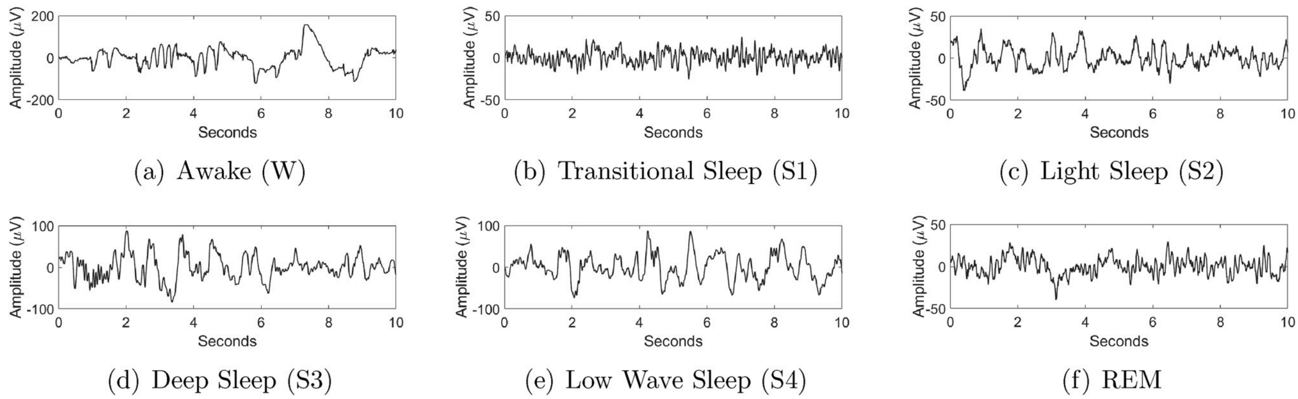

Fig. 3. Samples of the extracted sleep signals representing the six sleep stages: (a) Awake, (b) Transitional Sleep (S1), (c) Light Sleep (S2), (d) Deep Sleep (S3), (e) Low Wave Sleep (S4), (f) REM.

Table 1
Number and percentage of each sleep stage from the obtained EEG signals.

Sleep Stages	W	S1	S2	S3	S4	REM	Total
Total Number	72391	2805	17799	3370	2333	7717	106415
Percentage	68.03%	2.64%	16.73%	3.17%	2.19%	7.25%	100%

recordings of 20 h length each were obtained during day and night periods. The subjects wore a cassette-tape recorder and did not receive any sleep-related medications. On the other hand, the second data-set, named as *Sleep Telemetry (ST)*, includes 44 PSG recordings obtained for a study on the effect of temazepam on sleep in 1994. The subjects for the study are 22 Caucasian males and females that had difficulties getting asleep. Two 9-h PSG recordings were obtained from each patient in the hospital for two nights. All subjects wore a telemetry device to acquire the signals. Both PSG data-sets include EEG (Fpz-Cz/Pz-Oz channels), EOG, submental EMG, oro-nasal airflow, rectal body temperature, and event marker signals. The hypnograms included labels for all sleep stages (W, S1, S2, S3, S4, REM) scored manually by experienced technicians at every 30 s epochs based on the R&K scoring manual. Two additional scores, M and ?, are given to the data at any period representing movement and an undefined score, respectively. To provide sleep stages according to the current AASM scoring manual, stages S3 and S4 were manually merged into a single N3 stage.

For this study, only the raw EEG Fpz-Cz signals are utilized within the proposed algorithm. The selection of this channel was due to achieving better performance in sleep stages classification in several studies [61–63]. The recordings used were obtained from the SC data-set with a total of 39 signals obtained from 20 patients. One recording is missing due to cassette issues. The signals are sampled with a 100 Hz sampling rate without any prior filtering.

3.2. *Pre-processing*

To prepare the EEG signal for the classification procedure, the following steps are performed:

- EEG Signals Extraction.
- Signals Normalization.
- Savitzky-Golay Filtering.

Each step is briefly described in the following subsections.

3.2.1. *EEG Signals Extraction*

The data-set of EEG signals acquired from the PSG recordings described in subsection 3.1 contains continuous raw EEG signals corresponding to each patient. In addition, each signal comes with an annotation file that matches every 30 s epoch by its corresponding sleep stage score. Therefore, at each epoch, the signal is segmented and stored based on its sleep stage score. Fig. 3 shows samples of the six sleep stages signals extracted from the raw EEG signals. In addition, Table 1 shows the number and percentage of each sleep stage. The majority of the recordings included W signals with around 68.03%. Furthermore, S2 and REM had the highest amount of signals with 16.73% and 7.25%, respectively. It worth noting that after merging S3 and S4, the percentage of N3 stage following the AASM manual is 5.36%.

3.2.2. *Signals Normalization*

The extracted EEG signals are of large dynamic ranges in terms of signal amplitude. This causes smaller trends to be dominated by larger trends, and therefore, lose important information from the signals. To overcome this issue, signal normalization is applied to all sleep signals to ensure that they have a mean of zero and a unity variance ($\mu = 0, \sigma^2 = 1$). This step results in correctly identifying outliers within the sleep signals, which is considered an advantages in identifying differences between different signals. In literature, this step has been proven to improve the overall performance of training results in neural networks [64–66]. The normalization function selected is (`normalize()`) utilized from the MATLAB signal processing toolbox [67].

3.2.3. *Savitzky-Golay Filtering*

Savitzky-Golay (SG) Finite Impulse Response (FIR) filter is used to filter the EEG signal before the classification. SG is a digital filter that is used to smooth signals and increase its overall Signal-to-Noise (SNR) ratio [68]. The SG filter fits any input signals with a low order polynomial based on the linear least-square method. The least-square method results in having a set of polynomial coefficients. The convolution of all polynomial coefficients is applied to all the signal data points to estimate a smother signal [69,70].

If a signal has a set of data with a $n\{x_j, y_j\}$ points, where x is the independent variable and y is the corresponding observed value for all $j = 1, 2, \ldots, n$ points, the set of m polynomial coefficients, C_i, is used to obtain the smoothed signal, Y_j, as follows,

$$Y_j = \sum_{i=-(m-1)/2}^{i=(m-1)/2} C_i y_{j+1} \frac{m+1}{2} \tag{10}$$

In this study, the filter inputs used are the polynomial order and frame length. Based on trial and error, these values were selected to be 3 and 25, respectively. In addition [70], recommended the use of polynomial order of 2–4 and a frame length of 17–25 for a better SNR adaptive filtering. The filter is used within MATLAB with the function (`sgolayfilt()`). As features are derived automatically within the training model, SG filtering allows LSTM layers to easily extract features at each time step as there is less noise contaminated within signals.

3.3. Classification and training

Before starting with the classification and training process for the LSTM network, the filtered signals are split into two parts; training and testing signals. The splitting scenario is done using a k-fold cross-validation scheme of 10-folds. This scheme allows the training and testing over the whole dataset. Thus, the performance of the algorithm can be evaluated after going through all signals. In addition, two major steps are followed to ensure successful automatic sleep stage scoring; LSTM network layers selection and training parameters setting.

3.3.1. LSTM network layers

The LSTM network architecture is selected to be either an Uni-directional or Bi-directional LSTM to process the data in both the forward chain alone, and with both the forward and backward chains at the same time. The features of each signal are derived at every time step $x = \{x[0], x[1], x[2], \ldots, x[n]\}$ based on the selected training model. LSTM learns the features on time increasing manner if Uni-directional was selected. On the other hand, it learns features on both time sequences if Bi-directional was selected. In addition, the network consists of four major layers. These layers are the input sequence layer, fully connected, softmax, and classification layers. The input sequence layer is specified to have 3000 sequence data representing the total number of elements per signal. The fully connected layer provides the output sequence, specified to have 6 or 5 classes, by multiplying the input sequence by a weight matrix with the addition of a bias vector [60]. For the prediction of class labels, a softmax layer and a classification layer are added to the network.

3.3.2. Training parameters

The training parameters are selected to provide efficient LSTM training process. The number of nodes (hidden-units) in the LSTM model should be carefully selected to prevent over-fitting in the model during training [71]. The following equation is followed,

$$N_h = \frac{N_s}{\alpha \times (N_i + N_o)} \tag{11}$$

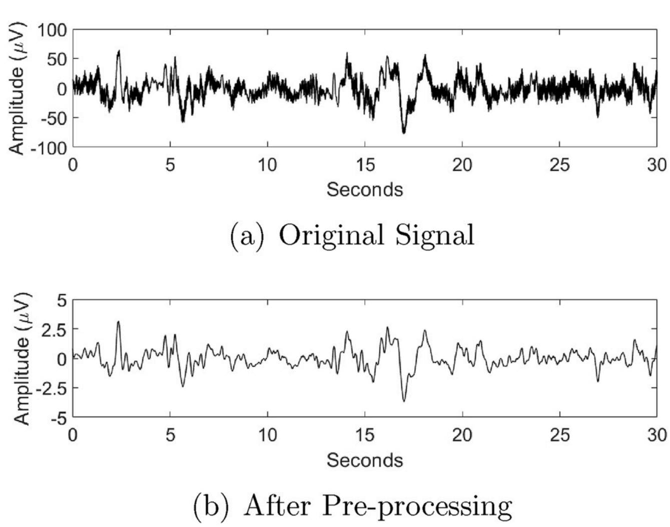

Fig. 4. Sample from the signals before and after z-score normalization and SG filtering: (a) Original Signals, (b) After Pre-processing.

where N_h is the number hidden nodes, N_s is the number of training samples, N_i is the number of input neurons, and N_o is the number of output neurons. The equation requires a multiplication by a scaling factor, α, which is usually a value between 2 and 10. In the current study, N_h is set to be 10 based on having N_i as 3000 for input nodes and N_o as 1 for each output class. In addition, the value of α is chosen to be the lowest (0.2) due to having around 50000 training samples. Furthermore, to evaluate the gradient of the loss function at each iteration during training, a sub-set of the training data, named as a mini-batch, is set to have a size of 50. The decision of having a small size of batches was to ensure quicker convergence and reduce the calculations complexity while training. It has been found in literature [72,73] that a default mini-batch size of 32 or 64 is recommended for better training performance. Therefore, a value in this range ensures a balance between having a good performance and requiring less computational complexity as well as time during training. It worth noting that the mini-batch size usually requires a trial and error selection approach based on the amount and type of the data used in training. The optimizer is chosen to be the Adaptive Moment Estimation (ADAM) solver. ADAM provides the ability to compute adaptive learning rates for the first and second moments of the gradients of different parameters [74]. The optimizer is set to have a learning rate (α) of 0.001, β_1 of 0.9, and β_2 of 0.999 by default settings. In addition, the network is allowed to pass 30 times through the training data with a total number of epochs of 30. The selected gradient clipping threshold is 1.

3.4. Performance evaluation

To evaluate the performance of the LSTM classifier, the corresponding confusion matrix of the classification process is obtained. The confusion matrix is a table that represents the classifier predictions for each sleep stage versus the manual experts' stage scoring (the gold standard). From the confusion matrix, several metrics are calculated including the accuracy, sensitivity, specificity, precision, and F1-score for each sleep stage. The F1-score represents an average between the sensitivity and precision [75]. The metrics are defined as follows,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{13}$$

$$Specificity = \frac{TN}{TN + FP} \tag{14}$$

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

$$F1 - score = \frac{2TP}{2TP + FP + FN} \tag{16}$$

where TP, TN, FP, and FN are the total number of true positives, true negatives, false positives, and false negatives, respectively. Furthermore, the definition of each parameter for each sleep stage based on the manual expert stage scoring is.

- True positives (TP): The total number of the sleep stage being classified correctly.
- True negatives (TN): The total number of other sleep stages being classified correctly as other sleep stages.
- False positives (FP): The total number of the sleep stage being classified incorrectly.
- False negatives (FN): The total number of other sleep stages being classified incorrectly as other sleep stages.

In addition, the Cohen's kappa (κ) coefficient, which is a value that

Table 2
Confusion matrix of the manual expert stage scoring and the Uni-directional LSTM automated stage scoring following the R&K manual.

		Expert Scoring						Total
		W	S1	S2	S3	S4	REM	
LSTM Scoring	*W*	**70896**	597	1227	500	410	346	73976
	S1	220	**1054**	350	21	5	123	1773
	S2	728	793	**15160**	1053	160	397	18291
	S3	158	19	498	**1359**	468	18	2520
	S4	130	4	83	394	**1277**	8	1896
	REM	259	338	481	43	13	**6825**	7959
Total		72391	2805	17799	3370	2333	7717	**106415**

Table 3
Confusion matrix of the manual expert stage scoring and the Bi-directional LSTM automated stage scoring following the R&K manual.

		Expert Scoring						Total
		W	S1	S2	S3	S4	REM	
LSTM Scoring	*W*	**70970**	434	914	335	218	340	73211
	S1	300	**1338**	422	14	7	151	2232
	S2	697	760	**15405**	922	117	421	18322
	S3	119	17	608	**1699**	443	20	2906
	S4	91	10	59	369	**1533**	5	2067
	REM	214	246	391	31	15	**6780**	7677
Total		72391	2805	17799	3370	2333	7717	**106415**

represents the agreement between two scoring decisions [76], is calculated for each sleep stage. This value is calculated as follows,

$$\kappa = \frac{P_0 - P_c}{1 - P_c} \tag{17}$$

where P_0 and P_c represent the observed agreements and the agreements expected by chance, respectively. In this study, the κ value represents the agreement between the LSTM classifier and the manual expert stage scoring, and is used to evaluate the overall algorithm performance.

4. Results and discussion

A sample from the signals before and after pre-processing is illustrated in Fig. 4. The z-score normalization and SG filtering makes it easier for the deep learning model based on LSTM to learn feature from signals. The averaging window was 25, this, a smoother signal was observed prior to training the models.

The algorithm was allowed to run on an Intel processor (i7-9700) with 32 GBs of RAM. The training process was performed on the NVIDIA GeForce GTX 1070 of 8 GBs display memory (VRAM). The duration of the training process was less than 2 h due to the cross-validation scheme followed in the current study. For a single fold, it took around 10 min to finish the process.

It worth noting that the data-set used included 106415 stages with the majority coming from the W stage. Unlike previous research works, the stages were not reduced and the LSTM models were trained using the exact data-set with no interference. Despite having imbalance in the stages, the model achieved acceptable levels of accuracy with no overfitting in the training phase.

4.1. Overall algorithm performance (R&K scoring manual)

The confusion matrix of the algorithm based on the R&K manual is shown in Table 2 and Table 3 for the Uni- and Bi-directional LSTM, respectively. The diagonal blocks represent the TP values of each sleep stage. From the table, these values are the highest in every row and column, which shows the success in classifying sleep stages. According to the confusion matrix of the Uni-directional LSTM, the overall number of TPs was 96571 with a 90.55% overall accuracy. On the other hand, the Bi-directional LSTM provided a total number of the TPs for all sleep stages of 97725, which is around 91.92% in accuracy relative to the total number of scores.

In addition, Table 4 and Table 5 shows the overall performance of the developed algorithm using the metrics described in subsection. 3.4 for each sleep stage using the Uni-directional and Bi-directional LSTM, respectively. For the Uni-directional LSTM, the algorithm achieved an average accuracy of 96.92% ± 1.50 with a Cohen's kappa value of 67.66% ± 20.59. The sensitivity and specificity of the system reached an average of 74.20% ± 16.45 and 97.91% ± 1.40, respectively. The precision ranged between 37.58% and 97.93% with an F1-score of 70.09% ± 22.11 on average. On the other hand, the Bi-directional algorithm achieved an accuracy of 94.12–98.62% for all the sleep stages. Furthermore, the specificity of the algorithm got an average value of 98.04% ± 1.33 between all sleep stages. The values of the sensitivity, precision, and F1-score is averaged at about 76.99% ± 15.61, 72.71% ± 21.14, and 74.64% ± 18.57. As for the Cohen's kappa value, the algorithm achieved a 72.55% ± 17.40 agreement between and developed LSTM algorithm and the manual expert stage scoring. It is clearly seen that the Bi-directional system slightly outperform the Uni-directional system, especially with the value of Cohen's kappa from 67.66% to

Table 4
The overall performance evaluation of the developed algorithm using Uni-directional LSTM in sleep stage scoring following the R&K manual.

Metrics	Cohen's Kappa (κ)	Accuracy	Sensitivity	Specificity	Precision	F1-score
W	89.99%	95.70%	95.84%	95.39%	97.93%	96.87%
S1	44.92%	97.68%	59.45%	98.33%	37.58%	46.05%
S2	80.75%	94.58%	82.88%	97.01%	85.17%	84.01%
S3	44.65%	97.02%	53.93%	98.06%	40.33%	46.15%
S4	59.60%	98.43%	67.35%	98.99%	54.74%	60.39%
REM	86.05%	98.10%	85.75%	99.09%	88.44%	87.08%
Average	**67.66%** ± **20.59**	**96.92%** ± **1.50**	**74.20%** ± **16.45**	**97.81%** ± **1.40**	**67.36%** ± **26.36**	**70.09%** ± **22.11**

Table 5
The overall performance evaluation of the developed algorithm using Bi-directional LSTM in sleep stage scoring following the R&K manual.

Metrics	Cohen's Kappa (κ)	Accuracy	Sensitivity	Specificity	Precision	F1-score
W	92.04%	96.56%	96.94%	95.72%	98.04%	97.48%
S1	52.01%	97.78%	59.95%	98.59%	47.70%	53.13%
S2	82.29%	95.01%	84.08%	97.28%	86.55%	85.30%
S3	52.76%	97.30%	58.47%	98.39%	50.42%	54.14%
S4	69.04%	98.75%	74.17%	99.23%	65.71%	69.68%
REM	87.16%	98.28%	88.32%	99.05%	87.86%	88.09%
Average	**72.55%** ± **17.40**	**97.28%** ± **1.35**	**76.99%** ± **15.61**	**98.04%** ± **1.33**	**72.71%** ± **21.14**	**74.64%** ± **18.57**

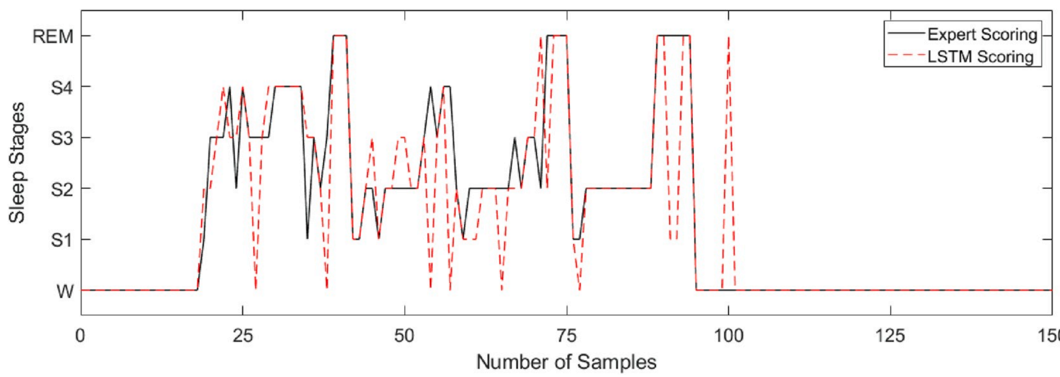

Fig. 5. A sample of the first 1000 classes showing sleep stage scoring of both the expert and the developed Bi-directional LSTM algorithm following the R&K manual.

Table 6
Confusion matrix of the manual expert stage scoring and the Bi-directional LSTM automated stage scoring following the AASM guidelines.

		Expert Scoring					Total
		W	N1	N2	N3	R	
LSTM Scoring	*W*	**70920**	497	974	409	324	73124
	N1	273	**1262**	366	20	140	2061
	N2	727	788	**15327**	948	451	18241
	N3	263	32	771	**4294**	20	5380
	R	208	226	361	32	**6782**	7609
Total		72391	2805	17799	5703	7717	**106415**

72.55%., which is expected as the Bi-directional system processes the data of each signal in both directions. Therefore, more information are learned in the training. Both methods provided good performance in the classification process, however, the Bi-directional LSTM model had higher values for the overall evaluation metrics. Despite having W stage with the highest number of stages, the model successfully identified other stages while training. This shows that other stages are identifiable compared with the W stages regardless of the number of stages included.

To elaborate more, a sample of the first 150 classes is illustrated in Fig. 5 showing both the manual expert stage scoring and the developed LSTM algorithm (Bi-directional). It can be seen from the figure that the algorithm achieved more than 90% of correct classification, which matches the observations of the confusion matrix shown in Table 3. The majority of the wrongly classified stages occurred on stages S1, S2, S3, S4, and REM. The W stage had very few wrongly classified stages compared to other stages for the whole classification process.

Furthermore, despite having the diagonal as the most predicted stages in the confusion matrices, it worth noting that the classifier is most likely predicting the preceding or the following stage. This is shown in Table 3 for stages W, S1, S2, S3, S4, and REM. Clinically, this could be correlated to the fact that there would be some kind of transition between stages [77], which makes the classifier uncertain towards predicting which stage is the correct one. For example, stage S3 and S4 are known to have a frequency range of 0–3 Hz. It is always difficult to differentiate those two stages clinically as they have no significant differences. In addition, they include sleep spindles and K-complexes of low amplitude waves. This is reflected in Table 5, as both stages are having almost the lowest κ measures. In addition, the transition is known to occur at the W to S1 and S2 to S3 with clear evidences. An S2 is considered as S3 or Slow Wave Sleep (SWS) if more than 20% of the epoch included delta waves activity. This may have took place in the data as shown in the predictions in Table 3. Furthermore, the classifier reads time-step features at every instance. This could results in having different observations from the decision made by manual expert scoring.

4.2. Comparison with AASM scoring manual

It worth mentioning that there is a drop in the sensitivity, precision, F1-score, and kappa values for both LSTM systems, which is mainly because of S1, S3, and S4 sleep stages. This could be due to the low number of stages in comparison to other sleep stages as observed from

Table 7
The overall performance evaluation of the developed algorithm using Bi-directional LSTM in sleep stage scoring following the AASM manual.

Metrics	Cohen's Kappa(κ)	Accuracy	Sensitivity	Specificity	Precision	F1-score
W	92.02%	96.55%	96.99%	95.58%	97.97%	97.47%
N1	50.77%	97.80%	61.23%	98.52%	44.99%	51.87%
N2	82.01%	94.94%	84.02%	97.30%	86.11%	85.06%
N3	76.25%	97.66%	79.81%	98.61%	75.29%	77.49%
R	87.61%	98.34%	89.13%	99.05%	87.88%	88.50%
Average	**77.73%** ± **16.20**	**97.06%** ± **1.35**	**82.24%** ± **13.35**	**97.85%** ± **1.41**	**78.45%** ± **20.36**	**80.08%** ± **17.33**

Table 8
Summary table of overall performance of this approach and other state-of-the-art deep learning algorithms.

Study	Year	Database	Total **Recordings**	EEG **Channel**	Total **Stages (W)**	Scoring **Method**	Extracted **Features**	Classification **Model**	Overall Performance			F1-score (Per-Stage)						Additional **Information**
									Accuracy	κ	F1-score	*W*	*S1*	*S2*	*S3*	*S4*	*REM*	
Langkvist et al. [78]	2012	St. Vincent's Recordings [79]	25	C3-A2	21,000 (4,767)	AASM	HMM-based Features	DBN	68.05%	63.00%	69.60%	73.00%	44.00%	65.00%	86.00%	–	80.00%	- S3 and S4 were considered together as one Slow-wave sleep (SWS or N3)
Tsinalis et al. [38]	2015	Sleep-EDF	39	Fpz-Cz	29,224 (3,380)	R&K	Morlet wavelets	SSA	74.80%	65.00%	69.80%	43.70%	80.60%	84.90%	74.50%	–	65.40%	- S3 and S4 were manually merged into N3 - Only In-bed recordings were considered - Fpz-Cz provided better performance
Tsinalis et al. [80]	2016	Sleep-EDF	39	Fpz-Cz	29,224 (3,380)	R&K	CNN-based Features	CNN	74.80%	71.00%	69.80%	72.00%	47.00%	85.00%	84.00%	–	81.00%	- S3 and S4 were manually merged into N3 - Only In-bed recordings were considered
Hassan et al. [81]	2017	Sleep-EDF	8	Pz-Oz	15,188 (8,055)	R&K	TQWT	Bootstrap Aggregating	92.43%	83.60%	–	–	–	–	–	–	–	- Sensitivity of each class: W = 98.61%, S1 = 37.42%, S2 = 92.93% S3 = 34.52%, S4 = 84.35%, REM = 82.34%
Hassan et al. [82]	2017	Sleep-EDF	8	Pz-Oz	15,188 (8,055)	R&K	EEMD	RUSBOOST	88.07%	88.00%	–	–	–	–	–	–	–	- Sensitivity of each class: W = 95.16%, S1 = 42.05%, S2 = 79.51% S3 = 86.61%, S4 = 48.09%, REM = 80.50%
Supratak et al. [35]	2017	Sleep-EDF	39	Fpz-Cz	15,188 (8,055)	R&K	CNN-based Features	CNN	82.00%	76.00%	76.90%	84.70%	46.60%	85.90%	84.80%	–	82.40%	- S3 and S4 were manually merged into N3 - Only 30 min before and after sleep
Supratak et al. [35]	2017	MASS [83]	62	F4-EOG (Left)	58,600 (6,227)	AASM			86.20%	80.00%	81.70%	87.30%	59.80%	90.30%	81.50%	–	89.30%	- S3 and S4 are considered together as one N3 stage according to AASM
Biswal et al. [84]	2017	Massachusetts General Hospital Recordings	10000	F3-M2, F4-M1 C3-M2, C4-M1	–	AASM	Spectrogram Features	RNN	85.76%	79.46%	80.60%	81.00%	70.00%	77.00%	83.00%	–	92.00%	- S3 and S4 are considered together as one

(*continued on next page*)

Table 8 (*continued*)

Study	Year	Database	Total **Recordings**	EEG **Channel**	Total **Stages (W)**	Scoring **Method**	Extracted **Features**	Classification **Model**	Overall Performance			F1-score (Per-Stage)						Additional **Information**
									Accuracy	κ	F1-score	*W*	*S1*	*S2*	*S3*	*S4*	*REM*	
																		N3 stage according to AASM
Li et al. [33]	2018	CCSHS [85]	116	O1-M2, O2-M1 C4-A2	130,046 (43,004)	R&K	Temporal Frequency Non-linear	RF	85.95%	80.46%	71.61%	91.30%	12.57%	87.11%	84.60%	–	82.46%	- S3 and S4 were manually merged into N3
Rahman et al. [59]	2018	Sleep-EDF	8	Horizontal EOG	15,188 (8,055)	R&K	DWT-based Features	RF SVM	90.38%	–	–	–	–	–	–	–	–	- Sensitivity of each class: W = 98.40%, S1 = 47.30%, S2 = 89.00% S3 = 55.00%, S4 = 81.00%, REM = 86.10%
Varela et al. [3]	2019	SHHS [86]	1000	C4-A2, C4-A1	1,209,971 (462,962)	AASM	CNN-based Features	CNN	–	83.00%	76.00%	95.00%	27.00%	88.00%	84.00%	–	86.00%	- S3 and S4 are considered together as one N3 stage according to AASM
Back et al. [12]	2019	Sleep-EDF	39	Fpz-Cz	42,308 (8,185)	R&K	RNN-based Features	IITNet	84.00%	78.00%	77.70%	87.90%	44.70%	88.00%	85.70%	–	82.10%	- S3 and S4 were manually merged into N3 - Only 30 min before and after sleep
Mousavi et al. [62]	2019	Sleep-EDF	39	Fpz-Cz	42,308 (8,285)	R&K	CNN-based Features	CNN	80.03%	73.00%	73.55%	91.72%	44.05%	82.49%	73.45%	–	76.06%	- S3 and S4 were manually merged into N3
This Study	**2020**	**Sleep-EDF**	**39**	**Fpz-Cz**	106,415 (72,391)	**R&K**	LSTM-based Features	BDLSTM	**97.28%**	**72.55%**	**74.64%**	**97.48%**	**53.13%**	**85.30%**	**54.14%**	**69.68%**	**88.09%**	- No removal of Wake stages was performed - Sensitivity of each class: W = 96.94%, S1 = 59.95%, S2 = 84.08% S3 = 58.47%, S4 = 74.17%, REM = 88.32%
						AASM			**97.062%**	**77.73..**	**80.08**	**97.47%**	**51.87%**	**85.06%**	**77.49%**	**-**	**88.50%**	- No removal of Wake stages was performed **- S3 and S4 were manually merged into a single N3 stage**

Table 1. In addition, according to the R&K scoring manual, both S3 and S4 are not merged into one stage, thus, the algorithm has to distinguish between their very minimal differences. The values of these metrics for the other sleep stages are almost more than 80%, which show a successful classification process. The complete confusion matrix following the AASM scoring manual is provided in Table 6. From the table, the overall TPs accuracy rate is 92.64%. Furthermore, Table 7 shows the overall performance evaluation of the algorithm using the Bi-directional LSTM (better performance) following the AASM manual. For the AASM calculations, both S3 and S4 were merged into a single N3 stage. This allowed the LSTM model to learn both stages as one stages, therefore, predictions of both stages became easier.

The value of Cohen's kappa (κ) increased from 72.55 $\pm$ 17.40 following the R&K to 77.73 $\pm$ 16.20 following the AASM manual is observed. In addition, the values of the sensitivity, precision, and F1-score are higher. The observations were expected due to the merging of S3 and S4 classes, which makes the classification process easier for the algorithm. In addition, The algorithm will no more learn to distinguish between S3 and S4, as they are considered from the same source (N3) following AASM.

Looking at stages N1, N2, and N3 from the confusion matrix, it is noted as well that there are more predictions toward the preceding sleep stage or the following. As previously discussed, similarities appear between signals at the transition step between sleep stages. This effect can be illustrated better due to the merging of S3 and S4 stages into a single N3. The majority of N2 predictions are toward the N3 stages, which could be correlated to having deep sleep (delta) waves contaminated within some epochs of the N2 stage [77]. Furthermore, this effect is not shown for stage R, as it is clinically easily distinguishable from other sleep stages by having higher activity in the brain unlike the relaxation at other stages.

4.3. Algorithm performance relative to state-of-the-art studies

The developed LSTM algorithm is reported with other State-of-the-Art studies on deep learning network for automatic sleep stage scoring. A total of 9 studies were covered during 2012–2019, with the utilization of different sleep signals databases such as the Sleep-EDF, St. Vincent's University Hospital and University College in Dublin, Montreal Archive of Sleep Studies (MASS), Massachusetts General Hospital, Cleveland Children's Sleep and Health Study (CCSHS), and Sleep Heart Health Study (SHHS) recordings. In addition, several feature extraction approaches were followed including Hidden Markov Model (HMM), Morlet wavelets, Tunable-Q Wavelet Transform (TQWT), Ensemble Empirical Mode Decomposition (EEMD), spectrogram, time, frequency, non-linear, and Discrete Wavelet Transform (DWT). Furthermore, classification models included Deep Belief Nets (DBNs), Stacked Sparse Autoencoder (SSAs), Convolutional Neural Networks (CNNs), Bootstrap aggregating, RUSBOOST, RNNs, Random Forests (RFs), Support Vector Machines (SVMs), and intra- and inter-epoch temporal context network (IITNet). The performance evaluation metrics are the overall accuracy, Cohen's kappa, F1-score, and per-stage F1-score. Table 8 summarizes all the observation of these studies in addition to the outcomes of the current study.

With regard to other studies, the current study achieved the highest value of accuracy using the Bi-directional LSTM system with 97.28%. The values of κ and F1-score were in the expected range of more than 72%. However, they were not the highest due to the merging of S3 and S4 stages in all other studies. When merging the two stages, the values improve for the Cohen's kappa and F1-score with 77.73% and 80.08%, respectively. Based on the Sleep-EDF database, 12 studies were included in the comparison between 2015 and 2019. This study reached the highest level of accuracy among other studies. In addition, the values of κ and F1-score were affected by the separation of S3 and S4 stages, which is not the case in other studies. However, it improved when following the same procedure of other studies by manually merging the two stages into a single N3 stage, represented as S3 in the table. It worth noting that the proposed classifier gave an overall sensitivity of almost 77.00% and 82.24% following R&K and AASM manuals, respectively. The low sensitivity and precision measures occurred mostly at the S1 (N1) and S3–S4 (N4) sleep stage detection. This could be correlated to the similarities between each sleep stage and the before/after sleep stages, as these stages are known to have relaxation in the overall brain response during sleep unlike the REM (R) sleep stage. The recent research works by Back et al. and Mousavi et al. used neural networks to train the model following the AASM scoring method. Their evaluation metrics were comparable to the performance of the proposed algorithm when merging stages S3 and S4. This could pave the way towards suggesting automatic model-based features as promising features to be learned for classification of sleep stages with no manual feature engineering.

5. Conclusion

In this paper, an investigation on the use of LSTM RNN in automatic sleep stage scoring is conducted. The developed algorithm assists in resolving the difficulties of manual sleep stage scoring done by experts at clinics. The algorithm successfully classified sleep stages to match the expert scoring with an overall 91.92% True Positive percentage using a Bi-directional LSTM system following the R&K manual. Relative to other studies, the algorithm is the highest in accuracy and within an acceptable range for both the Cohen's kappa and F1-score values. This study classified sleep stages based on the R&K scoring manual, which is challenging due to the minor differences between the S3 and S4 sleep stages. In addition, it showed an improved performance when following the recent AASM guidelines for sleep stage scoring.

The proposed algorithm can be implemented in various applications such as the detection of obstructive sleep apnea [87], epileptic seizures detection [88], and ECG arrhythmia detection [89]. In addition, at a patient-clinician level, it can be used to support clinicians decision making and overcome the difficulties in manual stage scoring including time consumption and human-errors.

The algorithm is usually able to detect time relation hidden inside the signals in both the forward and backward directions. A significant limitation for the LSTM algorithm is that it sees the inputs all at once. Despite the fact that this allows LSTM to outperform other machine learning algorithms in many tasks, it creates a huge load on the memory of the system and requires time to train the system [90].

The future work of this study includes further development on the LSTM network to enhance the overall performance. In addition, further testings on different databases requires to be performed relative to current data-set.

Compliance with ethical guidelines

The authors of the submitted Manuscript "*Sleep Stage Scoring using Long Short-Term Memory (LSTM) Learning System*" would like to declare that they have followed all the ethical guidelines in the institution and country where this work was conducted.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The author would like to thank the office of research at Abu Dhabi University for supporting this research.

References

[1] Tzimourta K, Tsilimbaris A, Tzioukalia K, Tzallas A, Tsipouras M, Astrakas L, Giannakeas N. EEG-based automatic sleep stage classification. Biomed J Sci Tech Res (BJSTR) 2018;7(4):6032–7.

[2] Hassan Ahnaf Rashik. Computer-aided obstructive sleep apnea detection using normal inverse Gaussian parameters and adaptive boosting. Biomed Signal Process Contr 2016;29:22–30.

[3] Fernández-Varela I, Hernández-Pereira E, Alvarez-Estevezy D, Moret-Bonillo V. A convolutional network for sleep stages classification. arXiv abs/1902.05748. arXiv:1902.05748. URL, http://arxiv.org/abs/1902.05748.

[4] Abdulla S, Diykh M, Laft RL, Saleh K, Deo RC. Sleep eeg signal analysis based on correlation graph similarity coupled with an ensemble extreme machine learning algorithm. Expert Syst Appl 2019;138:112790.

[5] Bousquet J, Khaltaev N. Global surveillance, prevention and control of chronic Respiratory diseases. World Health Organization (WHO); 2007.

[6] The American Academy of Sleep Medicine (Aasm). Insomnia factsheet, internet. URL, https://aasm.org/; 2008.

[7] Hassan AR, Haque MA. An expert system for automated identification of obstructive sleep apnea from single-lead ecg using random under sampling boosting. Neurocomputing 2017;235:122–30.

[8] Biswal S, Sun H, Goparaju B, Westover M, Sun J, Bianchi M. Expert-level sleep scoring with deep neural networks. J Am Med Inf Assoc 2018;25(12):1643–50.

[9] Lee K, Sheen S, Lee I, Choi B, Choi J, Park D, Kim H, Kim H. Scoring methods of polysomnography for diagnosis of sleep apnea in adolescents. Korean J Otorhinolaryngol-Head Neck Surg 2018;61(11):593–9.

[10] Nakamura T, Alqurashi YD, Morrell MJ, Mandic DP, Hearables. Automatic overnight sleep monitoring with standardized in-ear eeg sensor. IEEE (Inst Electr Electron Eng) Trans Biomed Eng 2019;67(1):203–12.

[11] Hassan Ahnaf Rashik, Kabir Muammar, Keshavarz Behrang, Taati Babak, Yadollahi Azadeh. Sigmoid wake probability model for high-resolution detection of drowsiness using electroencephalogram. In: 2019 41st annual international conference of the IEEE engineering in medicine and Biology society (EMBC). IEEE; 2019. p. 7080–3.

[12] Back S, Lee S, Seo H, Park D, Kim T, Lee K. Intra- and inter-epoch temporal context network (IITNet) for automatic sleep stage scoring. CoRR abs/1902.06562. arXiv: 1902.06562. URL, http://arxiv.org/abs/1902.06562.

[13] A. Rechtschaffen, A. Kales, A manual of standardized terminology and scoring system for sleep stages of human subjects, Public Health Service, US Government Printing Office, Washington DC.

[14] Malhotra R, Avidan A. Sleep stages and scoring technique. In: Chokroverty S, Thomas R, editors. Atlas of sleep medicine. Elsevier; 2014. p. 77–99.

[15] R. Berry, R. Brooks, C. Gamaldo, S. Harding, R. Lloyd, S. Quan, M. Troester, B. Vaughn, He AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications, American Academy of Sleep Medicine.

[16] Qureshi S, Kariila S, Vanichayobon S. Human sleep scoring based on K-bearest beighbors. Turk J Electr Eng Comput Sci 2018;26:2802–18.

[17] Fraiwan L, Khaswaneh N, Lweesy K. Automatic sleep stage scoring with wavelet packets based on single EEG recording. Int J Biomed Biol Eng 2009;3(6):85–8.

[18] Schwabedal J, Sippel D, Brandt M, Bialonski S. Automated classification of sleep stages and eeg artifacts in mice with deep learning. CoRR abs/; 1809, 08443.

[19] Zhou J, Wang G, Liu J, Wu D, Xu W, Wang Z, Ye J, Xia M, Hu Y, Tian Y. Automatic sleep stage classification with single channel eeg signal based on two-layer stacked ensemble model. IEEE Access 2020;8:57283–97.

[20] Berthomier C, Drouot X, Herman-Stoïca M, Berthomier P, Prado J, Bokar-Thire D, Benoit O, Mattout J, d'Ortho M. Automatic analysis of single-channel sleep EEG: validation in healthy individuals. Sleep 2007;30(11):1587–95.

[21] Jo H, Park J, Lee C, An S, Yoo S. Genetic fuzzy classifier for sleep stage identification. Comput Biol Med 2010;40(7):629–34.

[22] Koley B, Dey D. An ensemble system for automatic sleep stage classification using single channel EEG signal. Comput Biol Med 2012;42(12):1186–95.

[23] Huang C, Lin C, Ko L, Liu S, Su T, Lin C. Knowledge-based identification of sleep stages based on two forehead electroencephalogram channels. Front Neurosci 2014;8(263):1–12.

[24] Lajnef T, Chaibi S, Ruby P, Aguera P, Eichenlaub J, Samet M, Kachouri A, Jerbi K. Learning machines and sleeping brains: automatic sleep stage classification using decision-tree multi-class support vector machines. J Neurosci Methods 2015;250: 94–105.

[25] Ravan M, Begnaud J. Investigating the effect of short term responsive vns therapy on sleep quality using automatic sleep staging. IEEE (Inst Electr Electron Eng) Trans Biomed Eng 2019;66(12):3301–9.

[26] Diykh M, Li Y, Abdulla S. Eeg sleep stages identification based on weighted undirected complex networks. Comput Methods Progr Biomed 2020;184:105116.

[27] 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.

[28] Hassan A, Bhuiyan M. A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features. J Neurosci Methods 2016;271:107–18.

[29] Yulita I, Fanany M, Arymurthy A. Fast convolutional method for automatic sleep stage classification. Healthcare Inf Res 2018;24(3):170–8.

[30] Liang S, Huo C, Hu Y, Cheng Y. A rule-based automatic sleep staging method. J Neurosci Methods 2012;205(1):1–10.

[31] Hassan A, Bhuiyan M. Computer-aided sleep staging using complete ensemble empirical mode decomposition with adaptive noise and bootstrap aggregating. Biomed Signal Process Contr 2016;24:1–10.

[32] Sharma R, Pachori R, Upadhyay A. Automatic sleep stages classification based on iterative filtering of electroencephalogram signals. Neural Comput Appl 2017;28 (10):2959–78.

[33] Li X, Cui L, Tao S, Chen J, Zhang X, Zhang G. HyCLASSS: a hybrid classifier for automatic sleep stage scoring. IEEE J Biomed Health Inf 2018;22(2):375–85.

[34] Gunes S, Polat K, Yosunkaya S. Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting. Expert Syst Appl 2010;37(12):7922–8.

[35] 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.

[36] Sors A, Bonnet S, Mirek S, Vercueil L, Payen J. A convolutional neural network for sleep stage scoring from raw single-channel EEG. Biomed Signal Process Contr 2018;42:107–14.

[37] Chambon S, Galtier M, Arnal P, 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.

[38] Tsinalis O, Matthews P, Guo Y. Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders. Ann Biomed Eng 2016;44(5):1587–97.

[39] Yulita I, Fanany M, Arymuthy A. Bi-directional long short-term memory using quantized data of deep Belief networks for sleep stage classification. Procedia Comput Sci 2017;116:530–8. https://doi.org/10.1016/j.procs.2017.10.042.

[40] Zhao Y, Yang R, Chevalier G, Xu X, Zhang Z. Deep residual Bidir-LSTM for human activity recognition using wearable sensors. Math Probl Eng 2018:1. https://doi.org/10.1155/2018/7316954.

[41] Dong H, Supratak A, Pan W, Wu C, Matthews P, Guo Y. Mixed neural network approach for temporal sleep stage classification. IEEE Trans Neural Syst Rehabil Eng 2018;26(2):324–33.

[42] Phan H, Andreotti F, Cooray N, Chen O, Vos MD. Joint classification and prediction CNN framework for automatic sleep stage classification. IEEE (Inst Electr Electron Eng) Trans Biomed Eng 2019;66(5):1285–96.

[43] L. Ni and D. Wang and V. Singh and J. Wu and Y. Wang and Y. Tao and J. Zhang, Streamflow and rainfall forecasting by two long short-term memory-based models, J Hydrol:10.1016/j.jhydrol.2019.124296.

[44] Jozefowicz R, Zaremba W, Sutskever L. An empirical exploration of recurrent network architectures. In: The 32nd international conference on machine learning (ICML-15), lille, France; 2015.

[45] Cui Z, Ke R, Wang Y. Stacked bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction. CoRR abs/1801.02143. arXiv:1801.02143. URL, http://arxiv.org/abs/1801.02143.

[46] Sutskever I, Vinyals O, Le V. Sequence to sequence learning with neural networks. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, editors. Advances in neural information processing systems, vol. 27. Curran Associates, Inc.; 2014. p. 3104–12.

[47] Greff K, Srivastava R, Koutník J, Steunebrink B, Schmidhuber J. LSTM: a search space odyssey. CoRR abs/1503.04069. arXiv:1503.04069. URL, http://arxiv.org/abs/1503.04069.

[48] Malafeev A, Laptev D, Bauer S, Omlin X, Wierzbicka A, Wichniak A, Jernajczyk W, Riener R, Buhmann J, Achermann P. Automatic human sleep stage scoring using deep neural networks. Front Neurosci 2018;12:781.

[49] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8): 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735. URL.

[50] R. Fernandez, A. Rendel, B. Ramabhadran, R. Hoory, Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks, in: INTERSPEECH, 2014.

[51] Z. Cui, R. Ke, Y. Wang, Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction, arXiv preprint arXiv: 1801.02143.

[52] Graves A. Generating sequences with recurrent neural networks. CoRR abs/ 1308.0850. arXiv:1308.0850. URL, http://arxiv.org/abs/1308.0850.

[53] Jainand C, Medsker R. Recurrent neural networks: design and applications. first ed. Boca Raton, FL, USA: CRC Press, Inc.; 1999.

[54] Wang D, Nyberg E. A long short-term memory model for answer sentence selection in question answering. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing, vol. 2. Beijing, China: Association for Computational Linguistics; 2015. p. 707–12. https://doi.org/10.3115/v1/P15-2116. URL, https://www.aclweb.org/anthology/P15-2116.

[55] Schuster M, Paliwal K. Bidirectional recurrent neural networks. IEEE Trans Signal Process 1997;45(11):2673–81. https://doi.org/10.1109/78.650093.

[56] PhysioNet. Sleep-EDF database. URL, https://www.physionet.org/physiobank/database/sleep-edfx/; 2018.

[57] Aboalayon K, Faezipour M, Almuhammadi W, Moslehpour S. Sleep stage classification using EEG signal analysis: a comprehensive survey and new investigation. Entropy 2016;18:272.

[58] Hassan A, Bhuiyan M. Automatic sleep scoring using statistical features in the EMD domain and ensemble methods. Biocybernetics Biomed Eng 2016;36(1):248–55.

[59] Rahman Md Mosheyur, Bhuiyan Mohammed Imamul Hassan, Hassan Ahnaf Rashik. Sleep stage classification using single-channel EOG. Comput Biol Med 2018;102:211–20.

[60] Yildirim O, Baloglu B, Acharya R. A deep learning model for automated sleep stages classification using psg signals. Int J Environ Res Publ Health 2019;16(4): 599.

[61] Ghasemzadeh P, Kalbkhani H, Shayesteh MG. Sleep stages classification from eeg signal based on stockwell transform. IET Signal Process 2018;13(2):242–52.

[62] Mousavi S, Afghah F, Acharya A. SleepEEGNet: automated sleep stage scoring with sequence to sequence deep learning approach. PLOS ONE 2019;14(5):1–15. https://doi.org/10.1371/journal.pone.0216456.
[63] Jadhav P, Rajguru G, Datta D, Mukhopadhyay S. Automatic sleep stage classification using time–frequency images of cwt and transfer learning using convolution neural network. Biocybernetics Biomed Eng 2020;40(1):494–504.
[64] Aminian M, Aminian F, Ameli LSA. Electroencephalogram (EEG) signal classification using neural networks with wavelet packet analysis, principal component analysis and data normalization as preprocessors. In: Proceedings of the twenty-first MAICS 2010 midwest artificial intelligence; 2010.
[65] M. Alomari, A. Samaha, K. AlKamha, Automated classification of L/R hand movement EEG signals using advanced feature extraction and machine learning, arXiv preprint arXiv:1312.2877.
[66] R. Yannick, H. Banville, I. Albuquerque, A. Gramfort, T. Falk, J. Faubert, Deep learning-based electroencephalography analysis: a systematic review, arXiv preprint arXiv:1901.05498.
[67] MATLAB. Image processing toolbox. 2019. URL, https://www.mathworks.com/products/signal.html.
[68] Savitzky A, Golay M. Smoothing and differentiation of data by simplified least squares procedures. Anal Chem 1964;36(8):1627–39.
[69] Schafer R. What is a savitzky-golay filter. IEEE signal processing magazine 2011;28 (4):111–7.
[70] Acharya D, Rani A, Agarwal S, Singh V. Application of adaptive Savitzky–Golay filter for EEG signal processing. Perspect Sci 2016;8:677–9.
[71] M. Hagan, H. Demuth, M. Beale, O. De Jesús, Neural network design, second ed., Stillwater, Oklahoma. Oklahoma State University.
[72] Bengio Y. Practical recommendations for gradient-based training of deep architectures. In: Neural networks: tricks of the trade. Springer; 2012. p. 437–78.
[73] D. Masters, C. Luschi, Revisiting small batch training for deep neural networks, arXiv preprint arXiv:1804.07612.
[74] D. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980.
[75] Sasaki Y. The truth of the F-measure. Teach Tutor Mater 2007;1(5):1–5.
[76] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20 (1):37–46.
[77] Malhotra RK, Avidan AY. Sleep stages and scoring technique. Atlas of sleep medicine 2013:77–99.
[78] Längkvist M, Karlsson L, Loutfi A. Sleep stage classification using unsupervised feature learning. Adv Artificial Neural Syst 2012;2012:5.
[79] PhysioNet. St. Vincent's University hospital and University College Dublin sleep apnea database. URL, https://physionet.org/pn3/ucddb/; 2019.
[80] O. Tsinalis, P. Matthews, Y. Guo, S. Zafeiriou, Automatic sleep stage scoring with single-channel EEG using convolutional neural networks, arXiv preprint arXiv:1610.01683.
[81] Hassan Ahnaf Rashik, Subasi Abdulhamit. A decision support system for automated identification of sleep stages from single-channel EEG signals,. Knowl Base Syst 2017;128:115–24.
[82] Hassan Ahnaf Rashik, Bhuiyan Mohammed. Imamul Hassan, Automated identification of sleep states from EEG signals by means of ensemble empirical mode decomposition and random under sampling boosting. Comput Methods Progr Biomed 2017;140:201–10.
[83] O'reilly Christian, Gosselin Nadia, Carrier Julie, Nielsen Tore. Montreal Archive of Sleep Studies: an open-access resource for instrument benchmarking and exploratory research. J Sleep Res 2014;23(6):628–35.
[84] S. Biswal, J. Kulas, H. Sun, B. Goparaju, M. Westover, M. Bianchi, J. Sun, SLEEPNET: automated sleep staging system via deep learning, arXiv preprint arXiv:1707.08262.
[85] Resource NSR. Cleveland Children's sleep and health study (CCSHS). URL https://sleepdata.org/datasets/ccshs, ; 2019.
[86] National Heart, Lung, and Blood Institute. Sleep heart health study (SHHS). URL https://sleepdata.org/datasets/shhs, ; 2019.
[87] Liang X, Qiao X, Li Y. Obstructive sleep apnea detection using combination of cnn and lstm techniques. In: 2019 IEEE 8th joint international information Technology and artificial intelligence conference (ITAIC). IEEE; 2019. p. 1733–6.
[88] Nagabushanam P, George ST, Radha S. Eeg signal classification using lstm and improved neural network algorithms. Soft Comput 2019:1–23.
[89] Gao J, Zhang H, Lu P, Wang Z. An effective lstm recurrent network to detect arrhythmia on imbalanced ecg dataset. J Healthcare Eng 2019.
[90] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735–80.

Luay Fraiwan is a Professor of Biomedical Engineering currently working at Abu Dhabi University (ADU), on leave from Jordan University of Science and Technology (JUST). Prior to joining ADU, he was an Associate Professor of Biomedical Engineering at JUST. He was the Department Chair during the period 2011–2013. Dr. Luay obtained his Doctorate and Master's degrees in Biomedical Engineering in 2002 and 2005 respectively, from The University of Akron, Ohio, USA. Also, he holds a Master's degree in Medical Imaging from Compiegne Technological University, France. He got his Bachelor's degree in Electrical Engineering from JUST in 1993. Dr. Luay has a solid research experience in the field of biomedical signal processing and sleep analysis.

Mohanad Alkhodari received his Bachelor's degree with honor in Electrical Engineering from Abu Dhabi University (ADU), United Arab Emirates (UAE) in 2017. He received his Master's degree in Biomedical Engineering from American University of Sharjah (AUS), UAE in 2019. In summer 2017, he joined the Microsystems Lab (iMicro) at Masdar Institute, UAE, working on ultrasound systems design as a Research Assistant. While pursuing his Master's degree at AUS, he worked as a Teaching and Research Assistant in the College of Engineering under a full assistantship granted by the university to distinguished students. Alkhodari was a representative of the Electrical and Biomedical Engineering Master's programs for the IEEE Engineering in Medicine and Biology Society (EMBS) Student Chapter at AUS. His current research interests are deep learning, medical imaging, and biomedical signal processing.