Automatic Respiratory Event Scoring in Obstructive Sleep Apnea Using a Long Short-Term Memory Neural Network

Sami Nikkonen, Henri Korkalainen, Akseli Leino, Sami Myllymaa, Brett Duce, Timo Leppänen, and Juha Töyräs

***Abstract*—The diagnosis of obstructive sleep apnea is based on daytime symptoms and the frequency of respiratory events during the night. The respiratory events are scored manually from polysomnographic recordings, which is time-consuming and expensive. Therefore, automatic scoring methods could considerably improve the efficiency of sleep apnea diagnostics and release the resources currently needed for manual scoring to other areas of sleep medicine. In this study, we trained a long short-term memory neural network for automatic scoring of respiratory events using input signals from peripheral blood oxygen saturation, thermistor-airflow, nasal pressure -airflow, and thorax respiratory effort. The signals were extracted from 887 in-lab polysomnography recordings. 787 patients with suspected sleep apnea were used to train the neural network and 100 patients were used as an independent test set. The epoch-wise agreement between manual and automatic neural network scoring was high (88.9%, $\kappa = 0.728$). In addition, the apnea-hypopnea index (AHI) calculated from the automated scoring was close to the manually determined AHI with a mean absolute error of 3.0 events/hour and an intraclass correlation coefficient of 0.985. The neural network approach for automatic scoring of respiratory events achieved high accuracy and good agreement with manual scoring. The presented neural network could be used for analysis of large research datasets that are unfeasible to score manually, and has potential for clinical use in the future In addition, since the neural network scores individual respiratory events, the automatic scoring can be easily reviewed manually if desired.**

***Index Terms*—Machine learning, Artificial neural networks, Obstructive sleep apnea, Respiratory event scoring.**

Manuscript received August 21, 2020; revised November 13, 2020 and January 13, 2021; accepted March 5, 2021. Date of publication March 9, 2021; date of current version August 5, 2021. This work was supported by the Academy of Finland (313697, 323536), Research committee of the Kuopio University Hospital Catchment Area (5041781, 5041780, 5041797, 5041767, 5041794, 5041768), Instrumentarium Science Foundation, Research Foundation for Pulmonary Diseases, Foundation of the Finnish Anti-Tuberculosis Association, Respiratory Foundation of Kuopio Region, Päivikki and Sakari Sohlberg Foundation, Paulo Foundation, Business Finland, Finnish Cultural Foundation and Tampere Tuberculosis Foundation. *(Corresponding author: Sami Nikkonen.)*

Sami Nikkonen, Henri Korkalainen, Akseli Leino, Sami Myllymaa, and Timo Leppänen are with the Department of Applied Physics, University of Eastern Finland, 70211 Kuopio, Finland, and also with Diagnostic Imaging Center, Kuopio University Hospital, 70210 Kuopio, Finland (e-mail: nikkonen@uef.fi; henri.korkalainen@uef.fi; akseli.leino@uef.fi; sami.myllymaa@uef.fi; timo.leppanen@uef.fi).

Brett Duce is with Sleep Disorders Centre, Princess Alexandra Hospital, Brisbane, QLD 4102, Australia, and also with the Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, QLD 4102, Australia (e-mail: brett.duce@health.qld.gov.au).

Juha Töyräs is with the Department of Applied Physics, University of Eastern Finland, 70211 Kuopio, Finland and Diagnostic Imaging Center, Kuopio University Hospital, 70210 Kuopio, Finland, and also with School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, QLD 4072, Australia (e-mail: juha.toyras@uef.fi).

Digital Object Identifier 10.1109/JBHI.2021.3064694

I. Introduction

OBSTRUCTIVE sleep apnea (OSA) is a common breathing disorder where the upper airways collapse intermittently during sleep causing cessations in breathing [1]. These breathing cessations cause repeated hypoxia and sleep fragmentation which can lead to daytime sleepiness and depression [2], [3]. OSA is also associated with stroke and heart failure and increases the risk of traffic and workplace accidents [1], [4]–[6]. OSA has been estimated to affect nearly half of the adult population making it a major global health problem [7], [8].

OSA diagnosis is based on daytime symptoms and on the apnea-hypopnea index (AHI) *i.e.*, the number of apnea and hypopnea events per hour of sleep [9]. The gold standard to determine the AHI is to perform an in-lab polysomnography (PSG). In current clinical practice, respiratory events are scored by reviewing the recorded PSG signals and manually annotating the detected events. According to the American Academy of Sleep Medicine (AASM) scoring rules, an apnea is scored when the airflow signal drops ≥ 90 % from the reference level for at least 10 s and a hypopnea is scored, when the airflow signal drops ≥ 30 % from reference level for at least 10 s causing an arousal or at least a 3 % drop in blood oxygen saturation [10]. Since a single patient can have hundreds of respiratory events, manual scoring of PSG recordings is very time-consuming and expensive. Some devices and analysis software offer the possibility of automatic scoring of respiratory events, but the accuracy of these automatic scoring algorithms has been shown to be relatively poor compared to manual scoring with underestimation of the AHI by the automatic methods [11]–[15]. Therefore, there is a clear need to develop more advanced automated scoring methods.

Artificial neural network (ANN) methods have been shown to be powerful tools in medical signal analysis and have also

been used in sleep science for highly accurate automated sleep staging [16]–[18]. In addition, there is a large number of studies that use ANN-based methods to estimate sleep apnea severity using various input signals [19], [20]. However, these previous studies have only estimated the AHI or oxygen desaturation index (ODI), performed simple binary classification to OSA and non-OSA groups, or detected whether some signal segment includes respiratory events, but not their exact start time or duration [19], [20]. Therefore, these automatic methods cannot be directly compared to standard manual respiratory event scoring. Additionally, since these automatic methods do not provide any information on the individual respiratory events, they are effectively just "black boxes" that output an estimate of the OSA severity. Therefore, it is difficult to review the automatic analysis or confirm the severity estimate without completely re-scoring the signals manually.

For these reasons, the aim of this study was to develop an automatic respiratory event scoring method that allows the detection of individual respiratory event start times and durations. With this approach, the automatic scoring can be directly compared to standard manual scoring. In addition, the automatic scoring can be easily reviewed visually if desired. To achieve this, we used an ANN with a long short-term memory (LSTM) architecture that uses peripheral blood oxygen saturation (SpO_2), thermistor-airflow, nasal pressure –airflow, and thorax respiratory effort signals as inputs.

II. Methods

A. Subjects and Signals

The patient population consisted of 887 patients with suspected OSA who had undergone an in-lab PSG. The PSGs were conducted using the Compumedics Grael acquisition system (Compumedics, Abbotsford, Australia) during 2015-2017 in Princess Alexandra Hospital, Brisbane, Australia. The PSGs were analyzed by a group of ten expert scorers, with only a single person analyzing each recording, using the prevailing AASM guidelines (AASM 2012) [10], [21]. Ethical permissions for the data collection and processing were obtained from The Institutional Human Research Ethics Committee of the Princess Alexandra Hospital (HREC/16/QPAH/021 and LNR/2019/QMS/54313). The characteristics of the patient population are presented in Tables I and II.

All 887 recordings were included for further analyses. To keep the preprocessing steps simple and easily reproducible, we did not exclude any recordings for sleep duration or signal quality reasons and no artifact removal was performed. By including also the recording segments with poor signal quality to the current analyzed dataset, it is more representative of other clinical datasets which also generally lack preprocessing. Peripheral blood oxygen saturation, thermistor-airflow, nasal pressure –airflow, and respiratory effort signals were used as an input to the network. These signals were selected because they are the main signals utilized in manual respiratory event scoring [10]. Thermistor-airflow is included for accurate apnea detection and nasal pressure -airflow for accurate hypopnea detection [10]. In addition, these signals are recorded with most type II and type III portable monitors in addition to the type I in-lab PSGs [22], [23]. Therefore, a network trained with these four signals is applicable to almost any dataset. For this same reason, electroencephalography (EEG) recordings were not included as this would have limited the use in datasets recorded with type III monitors where EEG is not recorded [22].

TABLE I
Patient Characteristics for the Whole Dataset, Training Set and Test Set

		Mean	Standard deviation	Range
Whole dataset (n=887)	Age (years)	54.4	14.4	18.0-88.6
	AHI (events/hour)	28.3	28.1	0-146.1
	ODI 3% (events/hour)	19.2	28.5	0-145.1
	BMI (kg/m^2)	35.8	9.4	16.5-76.2
	ESS	10.0	5.6	0-24
Training set (n=787)	Age (years)	53.8	14.3	18.0-88.6
	AHI (events/hour)	27.9	27.7	0-146.1
	ODI 3% (events/hour)	18.9	28.1	0-145.1
	BMI (kg/m^2)	36.0	9.5	16.5-74.6
	ESS	10.0	5.6	0-24
Test set (n=100)	Age (years)	59.4	14.3	19.6-81.8
	AHI (events/hour)	29.6	25.8	0-100.7
	ODI 3% (events/hour)	21.2	31.3	0-127.6
	BMI (kg/m^2)	34.3	9.1	19.1-76.2
	ESS	9.7	5.4	0-21

AHI = apnea-hypopnea index, ODI = oxygen desaturation index, BMI = body mass index, ESS = Epworth sleepiness scale

TABLE II
The Number and Percentage of Patients and Respiratory Events in the Whole Dataset, Training Set and Test Set

		Number	Percentage
Whole dataset	Total number of patients	887	
	Male patients	489	55.1%
	Female patients	398	44.9%
	Total number of respiratory events	138 065	
	Number of apnea events	35 543	25.7%
	Number of hypopnea events	102 522	74.3%
Training set	Total number of patients	787	
	Male patients	430	54.6%
	Female patients	357	45.4%
	Total number of respiratory events	121 626	
	Number of apnea events	31 079	25.6%
	Number of hypopnea events	90 547	74.5%
Test set	Total number of patients	100	
	Male patients	59	59.0%
	Female patients	41	41.0%
	Total number of respiratory events	16 439	
	Number of apnea events	4 464	27.1%
	Number of hypopnea events	11 975	72.9%

The raw signals were imported to MATLAB 2019b (MathWorks Inc., Natick, Massachusetts, United States), for preprocessing. The thermistor, nasal pressure and thorax respiratory belt signals were originally recorded with 128 Hz sampling frequency and the oxygen saturation signal was recorded with 64 Hz sampling frequency. All four signals were lowpass filtered with a 2 Hz cutoff frequency and downsampled to 4 Hz. The downsampling was conducted to limit the computational load and the 4 Hz frequency was selected since nearly all of the power in the input signals is in the 0-2 Hz frequency range. Therefore, by selecting 4 Hz as the sampling frequency, no relevant signal information is lost and minimum computational load is achieved. In addition, the signals were truncated so that only the time between the lights off and lights on marks was included.

A scoring vector that contains the information of the manually scored apneas and hypopneas was formed for each patient. Each element in this scoring vector represents one data point in the input signals. The values of the scoring vector elements were set to one of three classes according to which respiratory event was annotated for that data point (0 = no-event, 1 = apnea, 2 = hypopnea). The apneas and hypopneas were not differentiated by event type (central, mixed or obstructive) and all types of events were included. These scoring vectors were used as the target outputs of the neural network. Therefore, the neural network effectively classifies each data point to no-event, apnea, or hypopnea. The input signals and the target vectors were split into 30 s epochs with 28 s overlap between consecutive epochs and each of these epochs was passed to the network as a single sample.

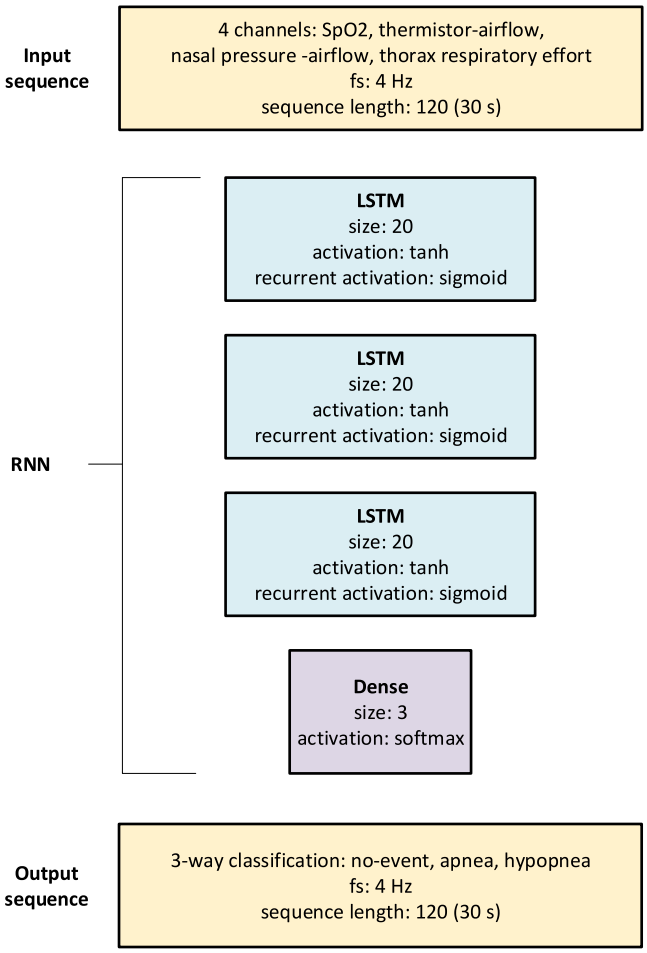

Fig. 1. Architecture of the neural network used in the study. LSTM = long short term memory layer, RNN = recurrent neural network, fs = sampling frequency.

B. Neural network

Recurrent neural network architecture was chosen since it allows sequence labeling, *i.e.*, classification of each sampling point. LSTM structure was selected since this type of network is well suited to process both long and short sequences while preserving relevant information throughout the sequence [24]. The network was trained in Python 3.7.3 with Tensorflow 1.14.0 using Keras 2.2.4. The training was conducted on a server with AMD Ryzen 2990WX, NVIDIA GeForce RTX 2080, and 128 GB RAM.

The neural network consisted of three LSTM layers, with a layer size of 20. The LSTM layers used tanh activation and sigmoid recurrent activation. The LSTM layers were followed by a fully connected layer with a size of 3 and a softmax activation. The network was trained with a learning rate of 0.0001 using the Adam optimizer [25]. An illustration of the neural network architecture is presented in Fig. 1.

The test set was formed by randomly selecting 100 patients from the full patient population. The rest of the patients (n = 787) were used to train the network. The patient characteristics of the training and test sets are also presented in Tables I and II. During training, 10% of the training set was further used as the validation set to assess the performance during training and to avoid overfitting. The training accuracy was monitored using sparse categorical cross-entropy as the loss function. The training was terminated after the validation set loss did not decrease for 200 consecutive epochs after which the model with the lowest validation loss was selected.

The neural network hyperparameters were selected based on educated guesses and preliminary testing. Networks with two and four LSTM layers were also tested but these networks resulted in lower validation accuracy. LSTM layer sizes of 10, 30, and 40 were also tested but the validation accuracy suffered slightly. The available training time and computational resources also limited the testing of larger networks and larger layer sizes. According to the preliminary testing, higher overlap in the input epochs improved accuracy slightly. However, 28 s of overlap on the epochs was the maximum allowed by memory constraints. Epoch lengths of 10 s, 20 s, 60 s, 120 s, and 300 s were also tested. On the epoch lengths of 60 s, 120 s, and 300 s, the percentage of overlap needed to be limited due to memory limitations, which likely led to the lower accuracy on the longer epoch lengths. On the shorter epoch lengths, the validation accuracy also suffered slightly compared to the 30 s epoch length likely due to the fact that the epoch length (10 s or 20s) was shorter than many of the respiratory events. Therefore, these hyperparameters were selected as according to the preliminary testing, they

enabled highest performance with the available training time and computational resources.

C. Data analyses

After training, the network performance was tested using the independent test set consisting of 100 patients. The neural network outputs from the patients in the test set also consisted of 30 s epochs with 28 s overlap. These outputs were separated by the event type into three scoring signals (no-event, apnea, and hypopnea). Each of the three signals was combined into a continuous signal by averaging the output values from the overlapping epochs. No other averaging or smoothing was performed on the outputs. These scoring signals therefore contained the predicted event probabilities for each data point. Since softmax activation was used on the output layer, the different event probabilities add up to one. The neural network scoring was formed by simply taking the maximum value from these output signals for each data point, *e.g.*, if apnea had a probability of 0.4 and hypopnea a probability of 0.5 and no event a probability of 0.1, that data point was scored as a hypopnea. Therefore, we obtained the neural network scoring vector with elements corresponding to the predicted event type (0 = no-event, 1 = apnea, 2 = hypopnea) for each data point without a need for any threshold values to qualify the events. The final neural network scoring was formed by combining consecutive data points with the same value to a single event. We considered that the event had ended at the first data point with a different output value. For example, apnea event started when the scoring vector value changed from 0 or 2 to 1 and ended when the value changed from 1 to 0 or 2. In accordance with the AASM rules, events with a duration of less than 10 s were discarded.

The agreement between the manual scoring and automated neural network-based scoring was evaluated epoch-wise. Each epoch was marked to have either an apnea, a hypopnea or no respiratory events. Partial events were also counted. Single epoch marked as containing an event could thus contain a part of an event, a single event, or multiple events. Epoch-wise accuracy and Cohen's kappa (κ) [26] were calculated for all respiratory events and separately for apnea and hypopnea scoring.

Additionally, the neural network scoring was compared to manual scoring in an event-by-event manner and the percentage of correctly detected events was calculated. The event was considered to be correctly detected if the neural network and manually scored events overlapped. Additionally, the errors in the event start and end times were calculated.

The AHI, apnea index (AI) and hypopnea index (HI) based on the neural network scoring were also calculated for each patient in the test set and compared to manual scoring. In addition, we calculated an intraclass correlation coefficient (ICC) [27] between the manually determined AHI, AI and HI and the AHI, AI and HI based on the neural network scoring.

III. Results

The neural network scoring had high agreement with manual scoring. The epoch-wise agreement for all respiratory events in the test set was 88.9 % ($\kappa = 0.728$). The epoch-wise agreement for apneas was 96.2 % ($\kappa = 0.791$) and the epoch-wise agreement for hypopneas was 88.2 % ($\kappa = 0.627$). The epoch-wise scoring for apnea, hypopnea and no-event epochs are presented in Table III. In addition, the sensitivity and specificity of the neural network scoring for apnea, hypopnea and no-event epochs are presented in Table III.

TABLE III
THE EPOCH-WISE SCORING RESULTS BASED ON MANUAL AND NEURAL NETWORK APPROACHES

	Number of epochs in manual scoring	Number of epochs in neural network scoring	Neural network scoring sensitivity	Neural network scoring specificity
Epochs scored as apnea	8 814	8 933	81.8%	97.8%
Epochs scored as hypopnea	18 808	15 433	63.7%	95.0%
Epochs scored as either respiratory event	26 635	22 904	74.8%	95.1%
Epochs scored as no-event	60 441	64 172	95.1%	74.8%

The total number of epochs in all recordings (n = 100) in the test set was 87 076. The same epoch can be scored both as an apnea and also as a hypopnea if it includes both events.

The AHI, AI, and HI calculated from the neural-network scoring were close to the manually determined AHI, AI, and HI. Mean absolute AI error was 2.0 events/hour, mean absolute HI error was 2.9 events/hour and mean absolute AHI error was 3.0 events/hour. Histograms showing the AHI, AI, and HI error distributions are presented in Fig. 2. The neural network estimated AHI, AI, and HI were highly correlated to manually determined AHI, AI, and HI and their scatter plots are also shown in Fig. 2. The intraclass correlation coefficient (ICC) between the neural network AHI and manual AHI was 0.985 with a 95 % confidence interval (CI) of 0.978 to 0.990. The ICC between neural network AI and manual AI was 0.971 (95 % CI: 0.955-0.981) and the ICC between neural network HI and manual HI was 0.966 (95 % CI: 0.943-0.979). Bland-Altman plots of the differences between the AHI, HI and AI obtained based on manual and neural network scoring are presented in Fig 3.

When the AHI based on neural network scoring was used to classify the test patients into the standard OSA severity groups, an overall accuracy of 87 % was achieved. Confusion matrix showing the OSA severity classifications based on manual and neural network scoring is presented in Fig 4. The errors in AHI, AI, and HI and classification accuracies were also calculated separately for each OSA class. These results are presented in Table IV. The neural network learning curves showing the training set and validation set performance are presented in Fig. 5.

In the event-by-event analysis, the neural network correctly detected majority of the respiratory events. Out of all manually

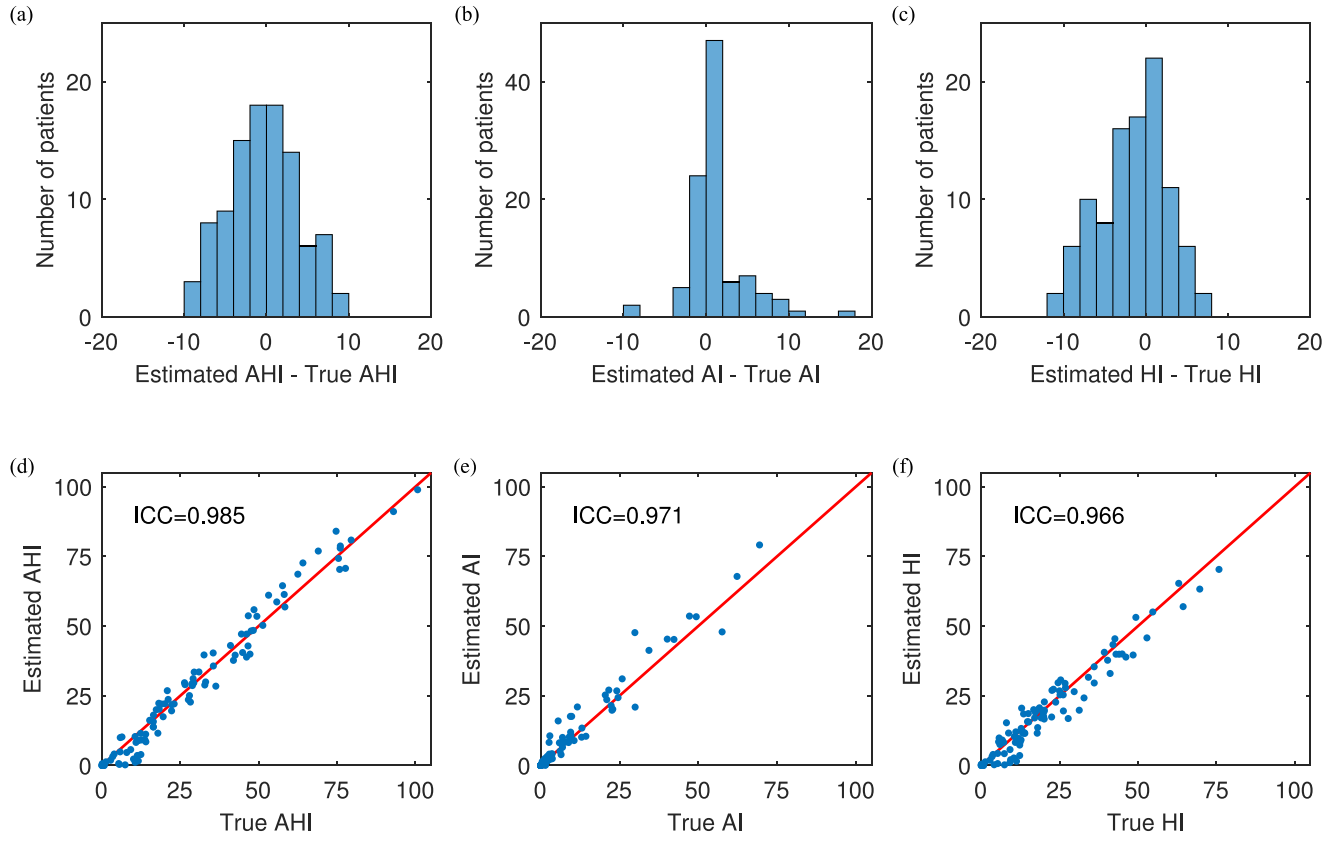

Fig. 2. Histograms of the differences between the neural network determined (estimated) and manually scored (true) apnea-hypopnea index (AHI) (a), apnea index (AI) (b) and hypopnea index (HI) (c) for all 100 patients in the test set. Scatter plots with intraclass correlation coefficients (ICC) between the estimated AHI and manually scored (true) AHI (d), between the estimated AI and true AI (e), and between the estimated HI and true HI (f) for all 100 patients in the test set.

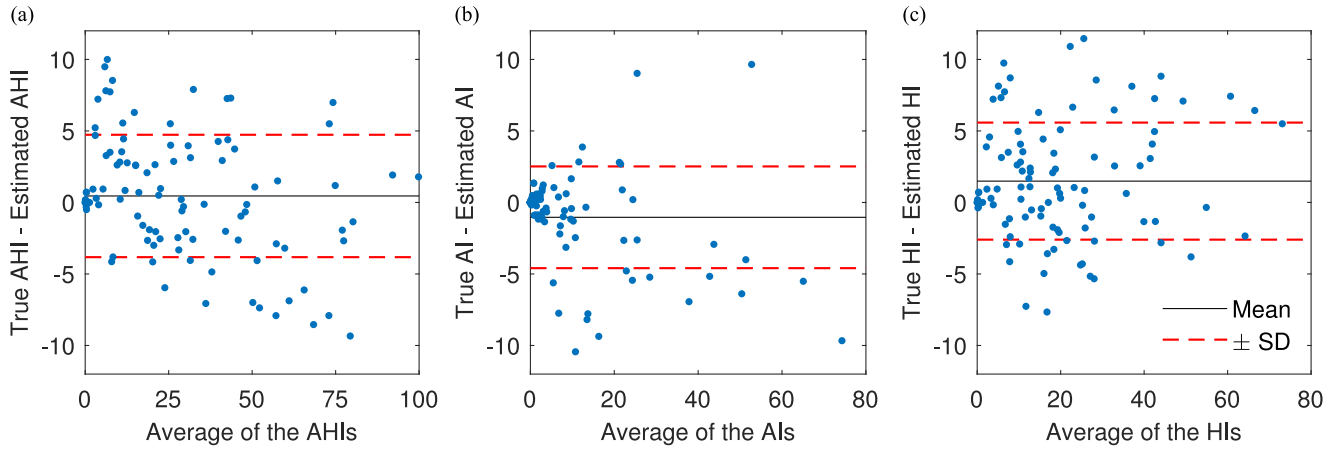

Fig. 3. Bland-Altman plots of the differences between the neural network determined (estimated) and manually scored (true) apnea-hypopnea index (AHI) (a), apnea index (AI) (b), and hypopnea index (HI) (c) for all 100 patients in the test set.

scored apnea events in the test set, 80.0 % were correctly detected by the neural network. For hypopneas, the percentage of correctly detected events was 60.1 %. The neural network also incorrectly identified some parts of the signals as respiratory events even though there was no manually scored event, *i.e.*, the events did not overlap with a manually scored event. Out of all apnea events scored by the neural network, 27.1 % were these misidentified events. For hypopneas, the misidentified event proportion was 34.2 %.

When the respiratory event was correctly detected by the neural network, there were still some differences in the event start and end points between the neural network scoring and

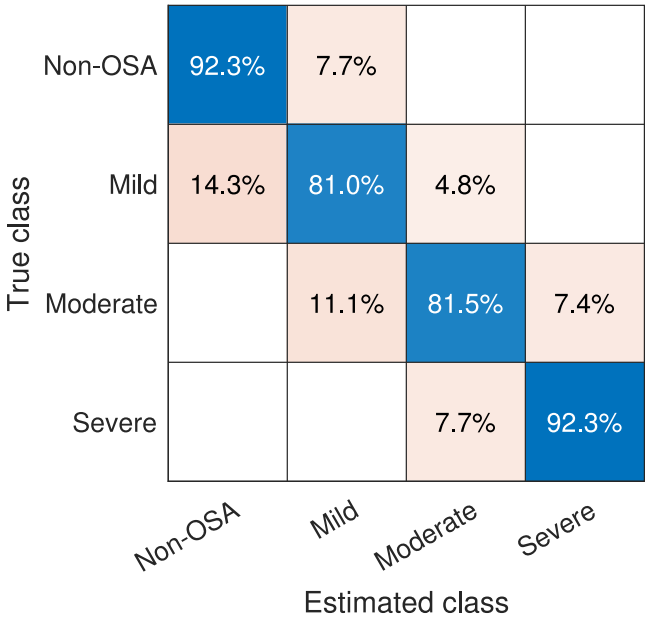

Fig. 4. Confusion matrix showing the obstructive sleep apnea (OSA) severity classifications based on manual scoring (true class) and neural network scoring (estimated class).

TABLE IV
THE AHI, AI AND HI ERRORS AND CLASSIFICATION ACCURACY FOR EACH OSA SEVERITY CLASS

	OSA severity based on manual scoring			
	Non-OSA n =13	Mild n =21	Moderate n =27	Severe n =39
Mean absolute AHI error (events/hour)	0.39	2.47	2.54	4.18
Mean absolute AI error (events/hour)	0.01	0.31	1.01	4.27
Mean absolute HI error (events/hour)	0.38	2.27	2.38	4.34
Classification accuracy (%)	92.3%	81.0%	81.5%	92.3%

OSA = obstructive sleep apnea, AHI = apnea-hypopnea index, AI = apnea index, HI = hypopnea index

manual scoring. The mean absolute error in the apnea start times across all correctly detected apneas in the test set was 6.6 s and the mean absolute error in end times was 1.9 s. Respectively, the mean absolute errors in hypopnea start and end times were 7.7 s and 4.4 s. The mean absolute errors in event durations were 7.2 s for apneas and 8.7 s for hypopneas. Histograms showing the distribution of the event start and end time errors and the event duration errors for apneas and hypopneas across all correctly detected events in the test set are presented in Fig. 6.

TABLE V
THE MEAN AND STANDARD DEVIATION OF THE ERROR PARAMETERS CALCULATED INDIVIDUALLY FOR EACH PATIENT IN THE TEST SET (N = 100)

	Mean	Standard deviation
Apnea detection %	78.3%	21.8%
Hypopnea detection %	54.7%	22.6%
Apnea misidentified event %	22.2%	24.7%
Hypopnea misidentified event %	35.3%	21.6%
Average absolute apnea start time error (s)	7.0	6.2
Average absolute hypopnea start time error (s)	6.7	3.9
Average absolute apnea end time error (s)	1.2	2.0
Average absolute hypopnea end time error (s)	2.6	2.2
Average absolute apnea duration error (s)	8.1	7.1
Average absolute hypopnea duration error (s)	8.9	5.0

The event detection percentages, the proportion of misidentified events, start time errors, end time errors, and duration errors were also calculated individually for each patient in the test set. These results are presented in Table V.

IV. DISCUSSION

In this study, we developed an artificial neural network that automatically scores respiratory events from sleep recordings. The epoch-wise agreement between manual and neural network scoring was high (88.9 %, $\kappa = 0.728$). The AASM inter-scorer reliability program reports a slightly better epoch-wise agreement for respiratory events (93.9 %, $\kappa = 0.92$) [28]. However, unlike in the present study, obstructive, central, and mixed apneas were scored separately in the AASM inter-scorer reliability program and the program investigated the agreement between more than 3000 scorers [28]. Furthermore, the records chosen to the AASM inter-scorer reliability program were specifically selected to only include robust signals with minimal artefacts [28]. For these reasons, the agreement values to the present study are not directly comparable. However, another study by Pittman *et al*. reports similar inter-scorer agreement (94.9 %, $\kappa = 0.82$) [29] as the AASM inter-scorer reliability program. This report also provides a more suitable comparison to the present study as it only included two manual scorers and did not differentiate between central and obstructive apnea types [29].

The presented neural network did not quite reach the scoring agreement reported by AASM and Pittman *et al*. [29]. However, as the neural network is trained using scoring data from multiple scorers, the maximum agreement that the network can achieve, is the agreement between the scorers of the training data. The neural network achieved respiratory event scoring agreement of $\kappa = 0.728$ with manual scoring which is close to the inter-scorer agreement of Princess Alexandra Hospital ($\kappa = 0.81$) [30].

The neural network estimated AHI, AI, and HI had high agreement with manual scoring (Fig 1). The mean absolute AHI error was 3.0 events/hour and the neural network achieved

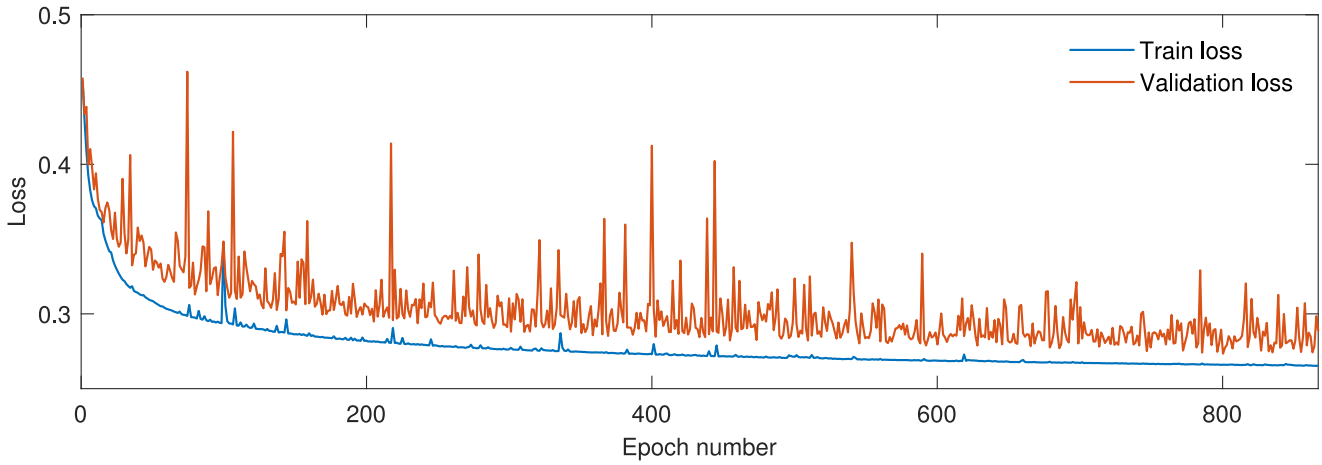

Fig. 5. The neural network learning curves showing the training set and validation set performance during training for each epoch.

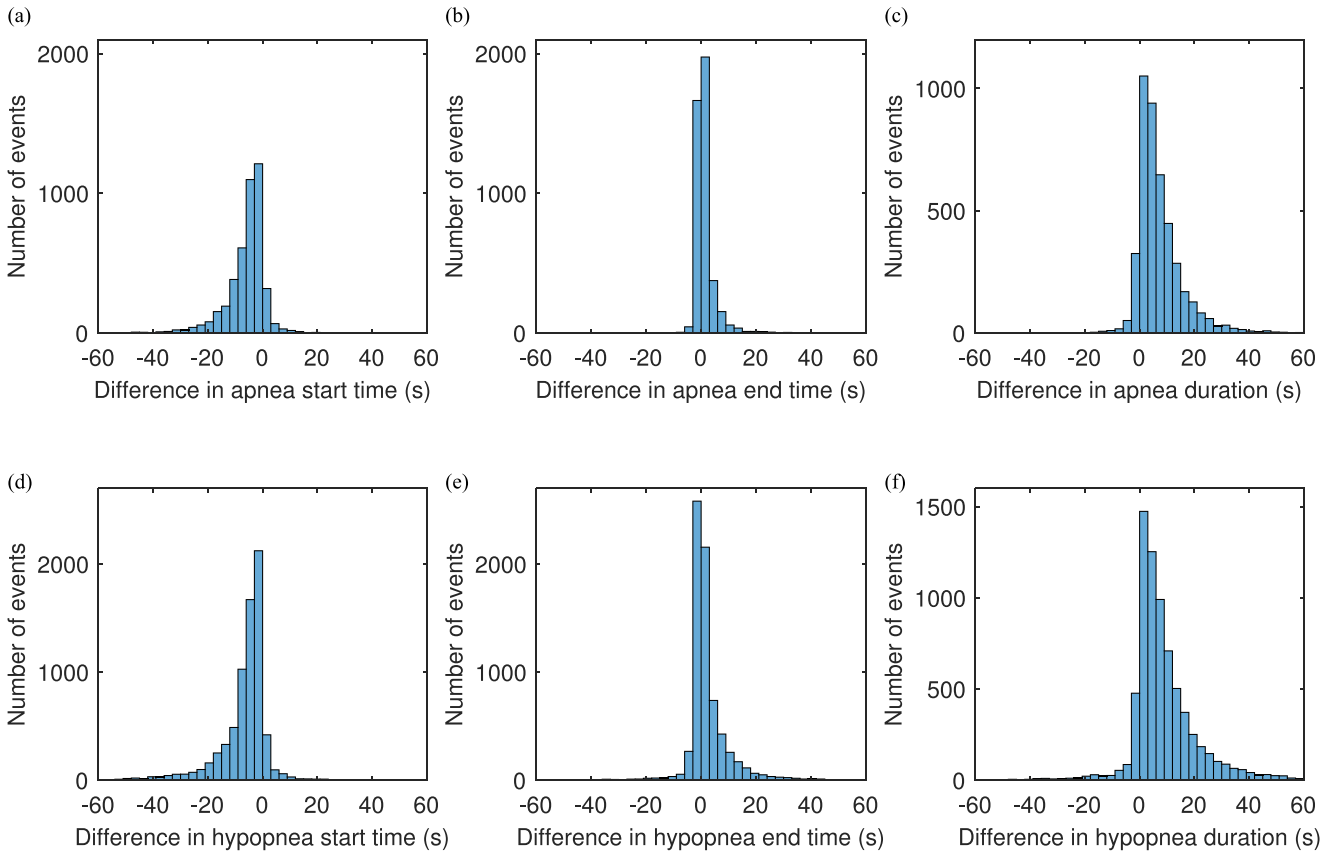

Fig. 6. Histograms of the apnea event start time errors (a), the apnea event end time errors (b), the apnea event duration errors (c), the hypopnea event start time errors (d), the hypopnea event end time errors (e), and the hypopnea event duration errors (f). All apneas and hypopneas in the test set are included. Start and end time errors are presented as manually scored event time - neural network scored event time, i.e., negative error means that the neural network scored event starts/ends later than the manually scored event and positive error means that the neural network scored event starts/ends before the manually scored event. Differences in event durations are presented as manually scored event duration - neural network scored event duration.

an AHI ICC of 0.985 with manual scoring. This is similar to reported inter-scorer agreement (AHI ICC of 0.95) among nine Sleep Apnea Genetics International Consortium (SAGIC) sleep centers [31].

The absolute AHI errors were higher in patients with more severe OSA (Table III). This was expected since the higher number of events also increases the probability of some events to be missed or misidentified. Classification accuracy was highest

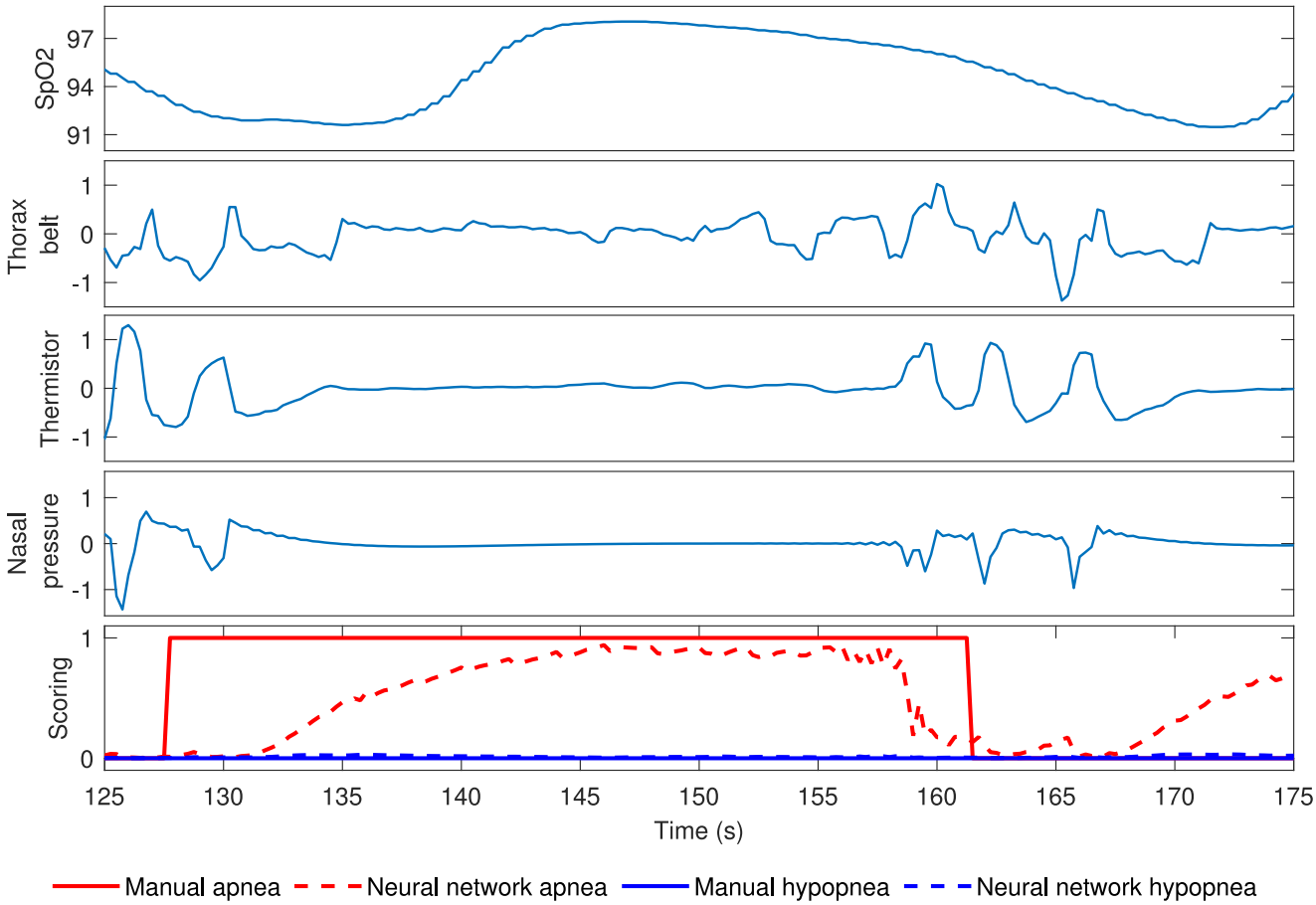

Fig. 7. Example of a respiratory event manually scored from breath peak to peak. In comparison, the neural network has only scored the event duration between the breaths. The neural network scoring is shown directly as the neural network outputs it, i.e., as a probability between 0 and 1. The closer the value is to 1, the higher the probability for an event there is according to the neural network estimation.

with non-OSA and severe patients (Table III). This was also expected since overestimating the AHI of a severe patient or underestimating the AHI of a non-OSA patient does not change the OSA class making these patients easier to classify. It is noteworthy that random chance has larger effect on these subclass errors and classification accuracies due to the relatively low number of patients in each class.

The validation set and training set losses during the training of the neural network decreased over time as expected although the validation loss was considerably more volatile than the training loss with relatively high peaks (Fig 5). This was also expected however, as the optimizer only works on the training set and therefore some adjustments to the weights of the network may increase the validation loss even though the training set loss decreases. The general trend of the validation set loss also decreased similarly to the training set trend.

The neural network correctly detected the majority of the respiratory events although the detection accuracy was considerably higher for apneas (80.0 %) than for hypopneas (60.1 %). Hypopneas were also misidentified more often (34.2 %) than apneas (27.1 %). Similar results were seen in the average absolute start and end time errors, which were lower for apneas (6.6 s and 1.9 s) than for hypopneas (7.7 s and 4.4 s). The higher accuracy for apneas was expected since apnea events are also easier to detect manually and hypopnea agreement between scorers has been reported to be much lower than apnea agreement [28]. In addition, since the neural network does not analyze EEG, it is possible that some hypopnea events that are only associated with an arousal, are missed. The lack of EEG could also cause misidentified events since it could be difficult to differentiate between spontaneous airflow amplitude drops and arousal associated hypopneas. Thus, it is possible that spontaneous amplitude drops in airflow signal are sometimes incorrectly identified as hypopneas by the neural network. These factors could limit the accuracy of hypopnea scoring and explain the lower accuracy for hypopneas.

The events scored by the neural network were generally shorter than the manually scored events as seen in Fig. 2. The neural network scored most apneas and hypopneas to start later and to end slightly earlier compared to manual scoring. This could be in part explained by the manual scoring habit of scoring from breath peak to peak while the neural network seems to only score the actual event duration. An example of this is presented in Fig. 7. Since the events scored by neural network were shorter, it could also at least partly explain the relatively high proportion of missed events. Many manually scored events with a duration

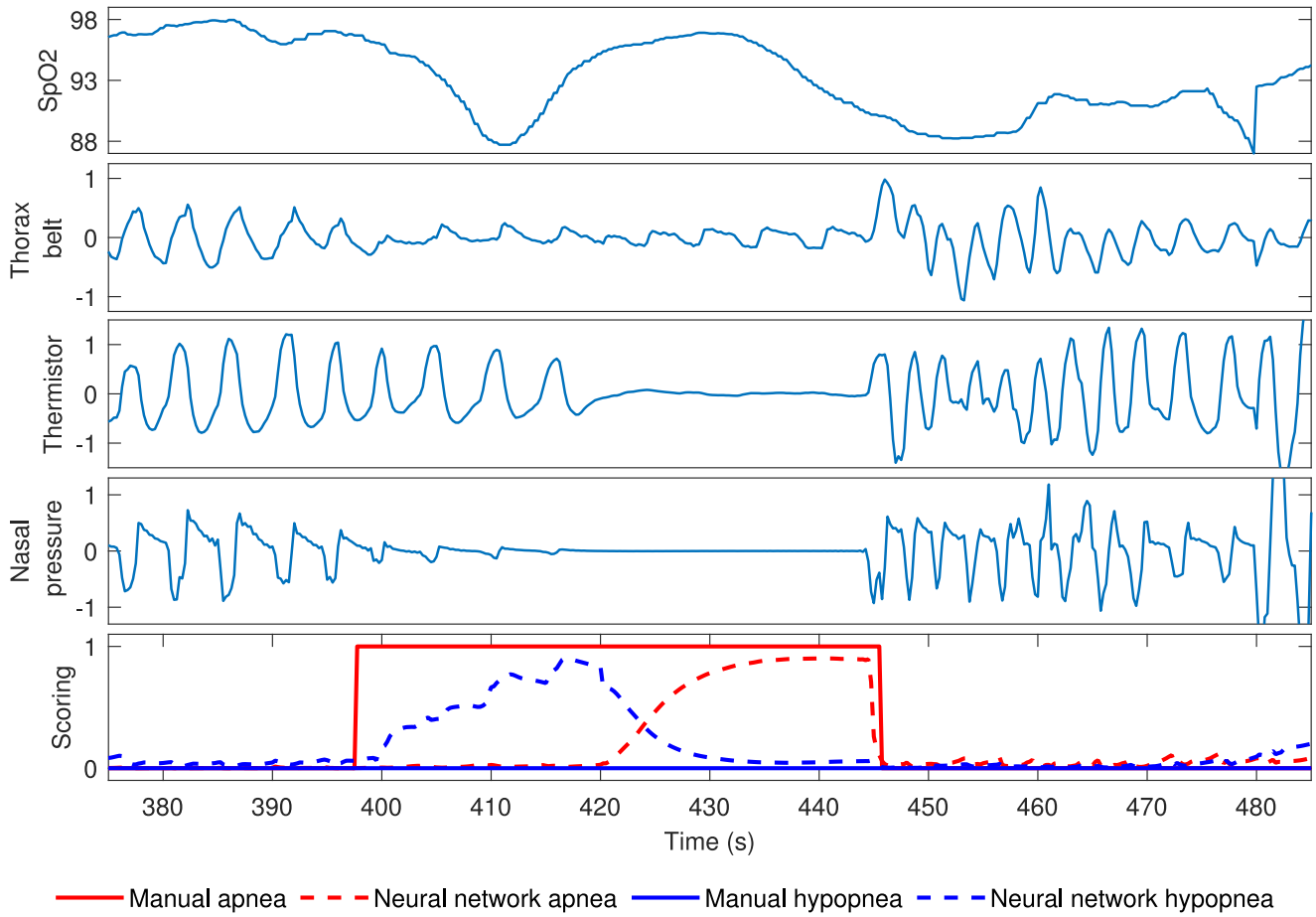

Fig. 8. Example of a respiratory event which starts as a hypopnea and turns into an apnea. The neural network has scored the event as separate hypopnea and apnea events while the whole event is scored as an apnea by manual scoring. The neural network scoring is shown directly as the neural network outputs it, i.e., as a probability between 0 and 1. The closer the value is to 1, the higher the probability for an event there is according to the neural network estimation.

close to 10 s were likely even shorter in the neural network scoring and may have thus been discarded since they did not fulfill the minimum duration criterion of 10 s.

While the neural network correctly detected majority of the events, it scored many of them differently than the manual scorers. In many cases, the neural network scores the beginning of the event as a hypopnea, when the thermistor still detected some airflow and then switched the event scoring to apnea after the thermistor signal was also flat. In contrast, the manual scorers typically scored the whole event as an apnea as instructed by the AASM apnea scoring rules [21]. An example of this is presented in Fig. 8. Some of the error in event start and end times could also be explained by this functional difference in human and neural network scoring. Therefore, the nuances within the scoring habits can have a significant impact on the accuracy of the neural network.

While the neural network -based scoring is not perfect, its accuracy is still relatively close to the inter-scorer agreement seen between manual scorers [28], [29], [31] and to the inter-scorer agreement of the training data [30]. In addition, the neural network scoring will not differ with different hospitals or scoring environments and always scores similar input the same way. This is a major factor as the accuracy of the manual scoring can be largely dependent on the training and scoring habits of the scorers. In some cases, the diagnosis for the same patient can even vary from healthy to severe OSA between different scorers [32]. In addition, the neural network is not susceptible to human error factors such as stress or alertness level making it very consistent. Furthermore, since the neural network produces scoring for all respiratory events, it can be reviewed visually should it be desired. The presented neural network is also fast, taking only a few seconds per patient, and does not require any manual labor to perform the scoring. For these reasons, the developed neural network solution could be applied for example in analyzing large datasets for research purposes where manual analysis of thousands of patients may be unfeasible due to time or cost constraints. Alternatively, the neural network could be used with portable monitors to perform and analyze sleep studies conducted at home. This way, sleep could be monitored for multiple consecutive nights quickly and cost-effectively and therefore the error caused by inter-night variation [33] of OSA severity could be reduced with minimal additional labor.

The presented neural network approach has certain limitations. One limitation is that we only used signals from pulse

oximetry (SpO_2), thermistor, nasal pressure sensor, and respiratory belt. As EEG signals were not included, hypopneas associated with only an arousal might not be detectable and therefore some of the hypopneas might be missed. However, the exclusion of EEG allows the neural network to be also utilized in datasets that do not include EEG. In addition, all of the used signals are easy to record and are present in all modern PSG and home sleep apnea test recording setups. Only some older datasets might not include a nasal pressure sensor and since all of these four input signals are required to use the neural network, this might limit the use of the neural network in these older datasets. In addition, the current neural network is not able to differentiate between obstructive, mixed, and central events. From a clinical point of view, this weakness could limit treatment decisions and therefore it warrants investigation in future studies whether this differentiation between central and obstructive apneas could also be added to the neural network. However, since the prevalence of central sleep apnea is very low [34], more training data, especially from patients suffering from central sleep apnea, would likely be also needed for accurate detection of central events. Another limitation is that the neural network has only been tested in the current dataset. While we have shown that the presented neural network has strong performance in a completely independent test set, no information is available on how the network performs in a completely different dataset which might have different manual scoring preferences and different recording setup. In addition, if different sensors such as different type of oximeter or respiratory belt are used, they might produce a slightly different signal which could affect the performance of the neural network. Therefore, it would be beneficial to evaluate in the future, how the neural network performs in another dataset. In addition, it would be interesting to investigate how the neural network responds to respiratory signal loss compared to manual scorers. Furthermore, it would be valuable to comprehensively study, which input signals are the most significant for the neural network performance and if the inputs could be further limited.

Finally, it is good to acknowledge that the epoch-wise agreement measure is only a rough estimate of scoring agreement and allows a relatively large variation in the scores while still retaining good agreement. Therefore, using this measure could hide some scoring errors by the neural network. However, as the epoch-wise agreement is the standard measure used in literature, we also used it in this study to allow comparison to the inter-scorer agreement in manual scoring. In addition, it should be noted, that when the error in AHI is used as a metric for studying scoring agreement, it allows some differences in the scorings while retaining high agreement in AHI. Since the neural network missed some events and misidentified others, these errors partially cancel each other out. However, the same effect is also present when comparing the scoring agreement between two manual scorers.

V. Conclusion

In conclusion, automatic, neural network -based scoring of respiratory events was found to be possible with high accuracy and high agreement with manual scoring. The presented automatic scoring method could be used to greatly reduce the work required for PSG scoring and enable diagnosis and treatment for many who are suffering from OSA, but are not diagnosed due to limited diagnostic resources. In addition, since the neural network provides easily interpretable scoring for individual respiratory events, the automatic scoring could be visually reviewed or corrected if a final manual check is preferred.

References

[1] A. S. Jordan, D. G. McSharry, and A. Malhotra, "Adult obstructive sleep apnoea," *Lancet*, vol. 383, no. 9918, pp. 736–747, 2014.

[2] C. Lal, C. Strange, and D. Bachman, "Neurocognitive impairment in obstructive sleep apnea," *Chest*, vol. 141, no. 6, pp. 1601–1610, 2012.

[3] L. S. Bennett, C. Barbour, B. Langford, J. R. Stradling, and R. J. O. Davies, "Health status in obstructive sleep apnea: Relationship with sleep fragmentation and daytime sleepiness, and effects of continuous positive airway pressure treatment," *Amer. J. Respir. Crit. Care Med.*, vol. 159, no. 6, pp. 1884–1890, 1999.

[4] T. Young, P. E. Peppard, and D. J. Gottlieb, "Epidemiology of obstructive sleep apnea: A population health perspective," *Amer. J. Respir. Crit. Care Med.*, vol. 165, no. 9, pp. 1217–1239, 2002.

[5] J. Teran-Santos, A. Jimenez-Gomez, and J. Cordero-Guevara, "The association between sleep apnea and the risk of traffic accidents," *New. Engl. J. Med.*, vol. 340, pp. 847–851, 1999.

[6] S. Horstmann, C. W. Hess, C. Bassetti, M. Gugger, and J. Mathis, "Sleepiness-Related accidents in sleep apnea patients," *Sleep*, vol. 23, no. 3, pp. 1–7, 2000.

[7] R. Heinzer *et al.*, "Prevalence of sleep-disordered breathing in the general population: The hypnolaus study," *Lancet Respir. Med.*, vol. 3, no. 4, pp. 310–318, 2015.

[8] A. Benjafield *et al.*, "Estimation of the global prevalence and burden of obstructive sleep apnoea: A literature-based analysis," *Lancet Respir. Med.*, vol. 7, no. 8, pp. 687–698, 2019.

[9] A. A. of Sleep medicine, "Sleep-Related breathing disorders in adults: Recommendations for syndrome definition and measurement techniques in clinical research," *Sleep*, vol. 22, no. 5, pp. 662–689, 1999.

[10] American Academy of Sleep Medicine, *AASM Manual for the Scoring of Sleep and Associated Events*. Darien, IL, USA: American Academy of Sleep Medicine, 2017.

[11] A. Bahammam, M. Sharif, D. E. Gacuan, and S. George, "Evaluation of the accuracy of manual and automatic scoring of a single airflow channel in patients with a high probability of obstructive sleep apnea," *Med. Sci. Monit.*, vol. 17, no. 2, pp. 13–19, 2011.

[12] R. N. Aurora, R. Swartz, and N. M. Punjabi, "Misclassification of OSA severity with automated scoring of home sleep recordings," *Chest*, vol. 147, no. 3, pp. 719–727, 2015.

[13] J. F. Masa *et al.*, "Effectiveness of sequential automaticmanual home respiratory polygraphy scoring," *Eur. Respir. J.*, vol. 41, no. 4, pp. 879–887, 2013.

[14] T. Verse, W. Pirsig, B. Junge-Hülsing, and B. Kroker, "Validation of the POLY-MESAM seven channel ambulatory recording unit," *Chest*, vol. 117, no. 6, pp. 1613–1618, 2000.

[15] J. M. Calleja, S. Esnaola, R. Rubio, and J. Durán, "Comparison of a cardiorespiratory device versus polysomnography for diagnosis of sleep apnoea," *Eur. Respir. J.*, vol. 20, no. 6, pp. 1505–1510, 2002.

[16] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," *Brief. Bioinform.*, vol. 19, no. 6, pp. 1236–1246, 2017.

[17] H. Sun *et al.*, "Large-scale automated sleep staging," *Sleep*, vol. 40, no. 10, 2017.

[18] H. Korkalainen *et al.*, "Accurate deep learning-based sleep staging in a clinical population with suspected obstructive sleep apnea," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 7, pp. 2073–2081, Dec. 2019.

[19] S. Nikkonen, I. O. Afara, T. Leppänen, and J. Töyräs, "Artificial neural network analysis of the oxygen saturation signal enables accurate diagnostics of sleep apnea," *Sci. Rep.*, vol. 9, pp. 1–9, 2019.

[20] M. B. Uddin, C. M. Chow, and S. W. Su, "Classification methods to detect sleep apnea in adults based on respiratory and oximetry signals: A systematic review," *Physiol. Meas.*, vol. 39, no. 3, 2018.

[21] A. A. of S. Medicine, "Rules for scoring respiratory events in sleep: Update of the 2007 AASM manual for the scoring of sleep and associated events," *J. Clin. Sleep Med.*, vol. 110, no. 3, pp. 249–254, 2012.
[22] R. Ferber *et al.*, "ASDA standards of practice: Portable recording in the assessment of obstructive sleep apnea," *Sleep*, vol. 17, no. 4, pp. 378–392, 1994.
[23] M. Ahmed, N. P. Patel, and I. Rosen, "Portable monitors in the diagnosis of obstructive sleep apnea," *Chest*, vol. 132, no. 5, pp. 1672–1677, 2007.
[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput,* vol. 9, no. 8, pp. 1735–1780, 1997.
[25] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations, ICLR*, 2015, pp. 1–15.
[26] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.
[27] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychol. Bull.*, vol. 86, no. 2, pp. 420–428, 1979.
[28] R. S. Rosenberg and S. Van Hout, "The american academy of sleep medicine inter-scorer reliability program: Respiratory events," *J. Clin. Sleep Med.*, vol. 10, no. 4, pp. 447–454, 2014.
[29] S. D. Pittman *et al.*, "Assessment of automated scoring of polysomnographic recordings in a population with suspected sleep-disordered breathing," *Sleep*, vol. 27, no. 7, pp. 1394–1403, 2004.
[30] T. Leppänen, A. Kulkas, A. Oksenberg, B. Duce, E. Mervaala, and J. Töyräs, "Differences in arousal probability and duration after apnea and hypopnea events in adult obstructive sleep apnea patients," *Physiol. Meas.*, vol. 39, no. 11, 2018.
[31] U. J. Magalang *et al.*, "Agreement in the scoring of respiratory events and sleep among international sleep centers," *Sleep*, vol. 36, no. 4, pp. 591–596, 2013.
[32] N. A. Collop, "Scoring variability between polysomnography technologists in different sleep laboratories," *Sleep Med,* vol. 3, no. 1, pp. 43–47, 2002.
[33] L. R. A. Bittencourt *et al.*, "The variability of the apnoea-hypopnoea index," *J. Sleep Res.*, vol. 10, no. 3, pp. 245–251, 2001.
[34] D. J. Eckert, A. S. Jordan, P. Merchia, and A. Malhotra, "Central sleep apnea: Pathophysiology and treatment," *Chest*, vol. 131, no. 2, pp. 595–607, 2007.