Automated Sleep Apnea Detection in Raw Respiratory Signals Using Long Short-Term Memory Neural Networks

Tom Van Steenkiste, Willemijn Groenendaal, Dirk Deschrijver, and Tom Dhaene

***Abstract*—Sleep apnea is one of the most common sleep disorders and the consequences of undiagnosed sleep apnea can be very severe, ranging from increased blood pressure to heart failure. However, many people are often unaware of their condition. The gold standard for diagnosing sleep apnea is an overnight polysomnography in a dedicated sleep laboratory. Yet, these tests are expensive and beds are limited as trained staff needs to analyze the entire recording. An automated detection method would allow a faster diagnosis and more patients to be analyzed. Most algorithms for automated sleep apnea detection use a set of human-engineered features, potentially missing important sleep apnea markers. In this paper, we present an algorithm based on state-of-the-art deep learning models for automatically extracting features and detecting sleep apnea events in respiratory signals. The algorithm is evaluated on the Sleep-Heart-Health-Study-1 dataset and provides per-epoch sensitivity and specificity scores comparable to the state of the art. Furthermore, when these predictions are mapped to the apnea–hypopnea index, a considerable improvement in per-patient scoring is achieved over conventional methods. This paper presents a powerful aid for trained staff to quickly diagnose sleep apnea.**

***Index Terms*—Sleep apnea, LSTM, deep learning, SHHS-1.**

I. Introduction

SLEEP apnea is one of the most common sleep disorders and is characterized by the occurrence of breathing pauses, also known as apneaic episodes, during the night which lead to frequent awakenings [1]. It is typically classified as either Obstructive Sleep Apnea (OSA) when the airway is blocked by the throat muscles, Central Sleep Apnea (CSA), when the signals to control the breathing are disturbed, or hypopnea, when the breathing becomes shallow. Hypopnea can further be categorized as either obstructive or central. Although some studies report that an estimated 49.7% of male and 23.4% of female adults suffer from sleep-disordered breathing [2], many cases remain undiagnosed as patients are rarely aware of their condition. These patients are at risk of hypertension, cardiac arrhythmia, heart attacks and strokes [3], [4]. Some studies also show that sleep apnea patients have an increased chance of being involved in motor vehicle collisions [5].

To diagnose sleep apnea, an overnight polysomnography (PSG) recording is performed in a specialized sleep laboratory [6]. During this PSG, multiple physiological signals, pertaining to respiration, oxygen saturation, cardiovascular functioning and sleep status are recorded. Afterwards, a trained sleep technician analyzes the data of the entire night and evaluates each part of the signal using a standard reference such as the American Academy of Sleep Medicine (AASM) guidelines [6] for the presence of sleep apnea. Each event in the signal is then annotated as either OSA, CSA or hypopnea. Often, only events that are clinically relevant (e.g., long apneas) are scored and shorter disturbances are unannotated. The annotations are summarized in an Apnea-Hypopnea-Index (AHI) which represents the number of apnea and hypopnea events per hour and which is used to categorize patients into a normal, mild, moderate or severe class.

As the amount of beds for PSG recording and the amount of trained sleep technicians for analysis are very limited, waiting times can get excessively long. These waiting times range in between 2 and 10 months in the UK, and in between 7 and 60 months in the USA [7]. Furthermore, high intra- and inter-scorer variability has been reported [8]–[10].

To increase the amount of people that can be analyzed, and to reduce these high intra- and inter-scorer variabilities, automated methods to assist the sleep technicians have been investigated. These methods range from rule-based algorithms to automated machine learning techniques and are generally based on human-engineered features. Determining which features to use and how many are needed to obtain the best predictive power, is a difficult task. Due to the human misinterpretations, potentially interesting sleep apnea markers in the biometric signals can be missed. Furthermore, noisy data can negatively impact the generalization properties of the models to new patients in practical settings.

In this work, a novel sleep apnea detection method is proposed, based on deep learning with long short-term memory

Manuscript received June 1, 2018; revised September 28, 2018 and November 13, 2018; accepted December 6, 2018. Date of publication December 10, 2018; date of current version November 6, 2019. This work was supported by imec. *(Corresponding author: Tom Van Steenkiste.)*

T. Van Steenkiste, D. Deschrijver, and T. Dhaene are with the imec, IDLab, Ghent University - imec, Gent B-9052, Belgium (e-mail: tomd.vansteenkiste@ugent.be; dirk.deschrijver@ugent.be; tom.dhaene@ugent.be).

W. Groenendaal is with the Holst Center/imec the Netherlands, Eindhoven 5656 AE, The Netherlands (e-mail: willemijn.groenendaal@imec-nl.nl).

Digital Object Identifier 10.1109/JBHI.2018.2886064

2168-2194 © 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

neural networks using raw physiological respiratory signals to automatically learn and extract relevant features, and detect potential sleep apnea events. The performance is compared to traditional human-engineered feature methods using data from the the Sleep-Heart-Health-Study-1 (SHHS-1) reference database [11], and the numerical results confirm that the proposed method outperforms the traditional methods when generalizing to noisy data from other patients measured in a real clinical setting.

Section II discusses related work in the field of machine learning for the automated detection of sleep apnea. Section III introduces the methodology of the new algorithm. Section IV explains the experimental setup and Section V demonstrates the results. Finally, in Section VI, conclusions are made.

II. Related Work

A. Physiological Signals

For accurate diagnosis of sleep apnea, trained staff use a variety of physiological signals. Over the years, many different sleep apnea detection methods have been proposed, based on a subset of these signals. Due to the intrinsic link between the respiratory system and sleep apnea, respiration and oximetry signals are commonly used [12]. Respiratory information can be extracted from nasal thermal sensors, pressure sensors near the mouth, conductive bands around the chest or other types of sensors. Oximetry measurements (typically SpO2) are also a valuable tool for diagnosing sleep apnea [13], [14], although by itself not sufficient [15].

The occurrence of sleep apnea is also reflected in other physiological signals, such as the electrocardiogram (ECG), which is typically heavily processed in order to extract relevant sleep apnea markers. An example of such pre-processing is performing heart rate variability analysis on ECG signals [16]. Another commonly used strategy is the extraction of respiratory information from ECG signals in a process known as ECG Derived Respiration (EDR) [17]. This is possible due to the respiratory motion being modulated on top of the ECG signal. However, other illnesses than sleep apnea can also significantly impact these signals. As these PSG measurements are uncomfortable for the patient, a lot of work has been done towards the development of portable monitors with less obtrusive sensors. An example of this is ballistocardiography for the detection of sleep apnea [18], [19].

B. Sleep Apnea Detection

When analyzing sleep, all types of sleep apnea have to be detected. Additionally, to get a complete assessment of sleep quality, other events such as teeth grinding and snoring also have to be detected. This is demonstrated by the large interest in the recent CinC challenge [20]. However, the focus of this work is on the detection of sleep apnea and as such, only the detection of sleep apnea events will be analyzed in this study.

Various algorithms have been developed for automatically detecting sleep apnea events in one or more of the physiological signals originating from an overnight PSG. A common approach is to use interpretable rule-based algorithms that provide a clear explanation as to why some epochs of the signal are flagged as containing a sleep apnea event or not [21]. In medicine, such white-box approaches are very valuable.

However, other approaches with a higher learning capacity, based on machine learning, can automatically detect more complex patterns and make more accurate predictions. Commonly used methods include Support Vector Machines (SVM) [22], Logistic Regression (LR) [23], K-Nearest-Neighbors (KNN) [22], [23], Linear Discriminant Analysis [23], [24], Gaussian Processes (GP) [25] and Artificial Neural Networks (ANN) [26]. These methods typically start with computing a set of human-engineered features over a certain epoch of the data. For each epoch, a prediction is made whether or not it might contain an apnea event. These methods do not capture the temporal correlation components that are present in physiological signals. Specialized models, such as the SVM-based discriminative Hidden Markov Model (HMM) [27] utilize this time information to improve the accuracy of the estimates. Recent developments in deep learning have led to another temporal sleep apnea detection model. Long Short-Term Memory (LSTM) neural networks, a type of Recurrent Neural Network (RNN), were proposed as a good method capable of detecting long-term as well as short-term correlations in time-series of human-engineered features for sleep apnea [28]–[30] as well as for other medical use-cases [31], [32]. Although such models have incorporated valuable information by integrating the time-based component, other valuable aspects of the data are still lost, due the need for human-engineered features that summarize the data into distinct values. Additionally, some of these models are trained and analyzed on a human-selected set of clean epochs. In practice, this leads to generalization issues when analyzing data of new patients in real noisy settings.

III. Proposed Algorithm

In this work, a novel method is proposed using the LSTM model. Instead of extracting human-engineered features, the models are trained using a noise-filtered version of the actual respiratory signal itself. The main goal of the algorithm is to provide as much information as possible to the deep learning network, such that it can automatically extract relevant respiratory markers for the detection of apnea events without the need for human feature engineering. The complete workflow of the training algorithm is shown in Fig. 1. The setup of this workflow is that of a typical machine learning process. The process starts with collecting respiratory data which is then pre-processed and split into separate epochs. Each of the epochs is annotated based on the annotation of trained sleep technicians. Then, the epochs are used in a balanced bootstrapping scheme to create separate datasets. Finally, each of the datasets is used to train a separate LSTM model. During prediction, an aggregation step is added. Each of these steps will be discussed in detail in the following subsections.

A. Data Collection

The first step in the algorithm is the collection of respiratory data. This can be extracted from various sources such as respiratory bands or the ECG. To train the models, the respiration data

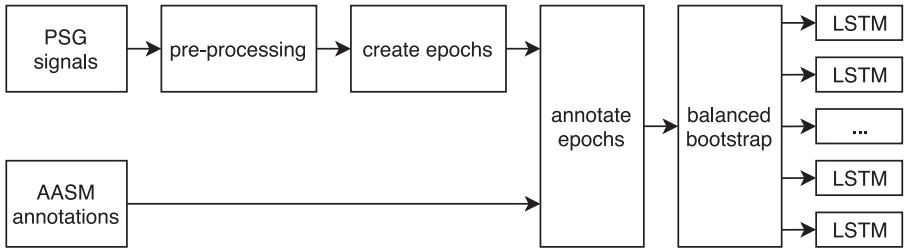


Fig. 1. Modeling sleep apnea: respiratory signals are pre-processed and combined with their label. Next, a balanced bootstrapping procedure combines epochs into datasets for training multiple LSTM networks.

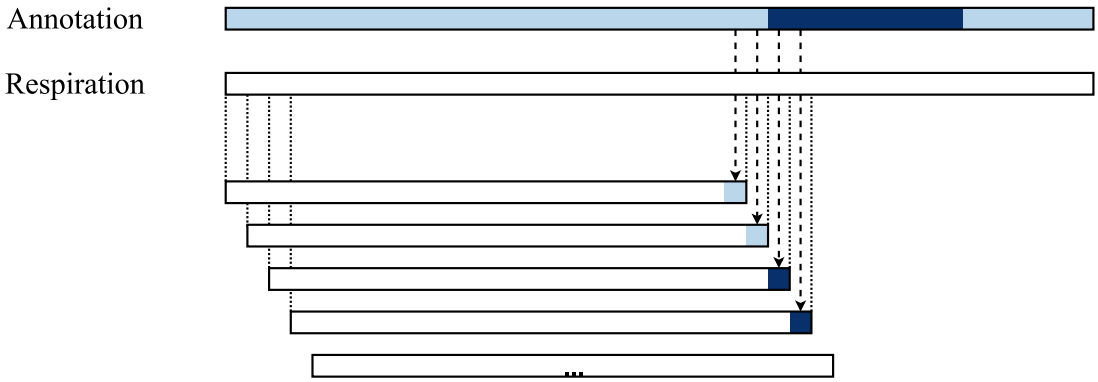


Fig. 2. The binary label of an individual epoch is determined based on the annotation of trained staff at the end of that epoch. The darker shaded portion indicates a sleep apnea event.

is combined with annotations. To create a robust model, the data must include a sufficient amount of patients with large enough variety in e.g., age, gender and physiology for which data of an entire night has been recorded.

B. Pre-Processing and Epoch Creation

Raw physiological signals contain a wide range of noise due to subject movement, electrical interference, measurement noise and other disturbances. In any sleep apnea detection method, noise canceling methods are essential and frequently used.

To extract relevant respiratory information, and to reduce noise, the physiological respiratory signals are passed through a fourth-order low-pass zero-phase-shift Butterworth filter with a cut-off frequency of 0.7 Hz [33]. This cutoff frequency is chosen to retain the major respiration components while removing as much noise as possible [34]. Next, motion artifacts of the patient and baseline wander are removed by subtracting a moving average filtered signal with a width of 4 seconds from the original signal. Finally, the sampling rate of the physiological signal is reduced to 5 Hz in order to speed up the analysis while still keeping the most relevant respiratory information. The mild filtering ensures that as much information as possible is kept in the signal such that it can be considered raw. Note that noisy sections of the signal are not removed from the dataset to accurately reflect real clinical settings.

The filtered signals are segmented into 30 second epochs with a stride of 1 second between them. Hence, the epochs are overlapping and each second of the data is represented in multiple epochs. Overlapping the epochs is not a typical strategy in sleep apnea research, but it offers some advantages. It allows the model to make predictions on a per-second basis, increasing the granularity of the detection. In addition, it significantly increases the amount of data that can be used for training the neural network.

C. Data Annotation

Finally, each of the epochs is labeled with annotations provided by a trained sleep technician that analyzed the data signals according to specific guidelines, such as the AASM [6] or the SHHS [11] guidelines. If at the end of an epoch, the sleep technician indicated an apneaic episode, the entire epoch was flagged as a positive apnea episode. This process is illustrated in Fig. 2.

Since the goal of this study is to detect all apneas and to provide a metric of the severity of apnea for each patient, and hence to provide a metric of the AHI, we combine all annotations into a single binary annotation (apnea or non-apnea).

D. Balanced Bootstrapping

Although sleep apnea is a common disorder, apnea-positive epochs are a relatively rare occurance for each patient. The majority of epochs is apnea-negative and only a small minority is apnea-positive. When such an imbalanced dataset is used to train machine learning models, most of these models will be heavily biased towards the majority class which may provide skewed results.

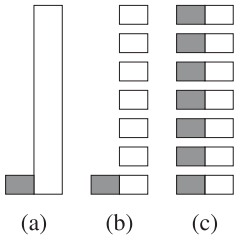


Fig. 3. The balanced bootstrapping procedure is used to transform an unbalanced dataset (Fig. 3a) into multiple balanced datasets. The majority class is divided into sub-datasets with size equal to the minority class (Fig. 3b). Subsequently, the minority class is copied to each of the sub-datasets to create balanced sub-datasets (Fig. 3c). The minority class is represented by the shaded bars while the majority class is represented by the clear bars.

There are several possibilities to cope with an imbalanced dataset. The most straightforward solution is to downsample the dataset [35], [36]. With downsampling, apnea-negative epochs are removed from the dataset until there are as many apnea-positive epochs as there are apnea-negative epochs. In practice, when dealing with a large data imbalance, this means a majority of the data is removed from the dataset and a lot of valuable information is lost.

Another commonly used method is oversampling the dataset [35], [36]. Apnea-positive epochs are duplicated in the dataset until there are as many apnea-positive epochs as apnea-negative epochs. Although all information is retained, this duplication increases the risk of overfitting to a subset of apnea-positive examples, heavily impairing the generalization power to new data.

To overcome these disadvantages of both methods, an innovative procedure called balanced bootstrapping has been proposed [37]. In this work, balanced bootstrapping is applied but instead of picking random samples, the entire minority class is used each time, as illustrated in Fig. 3. The large imbalanced dataset is split up into several smaller balanced datasets. First, the majority class in the unbalanced dataset is split into subsets with size equal to the minority class. Then, the epochs of the minority class are appended to the different sets of the majority class leading to multiple balanced datasets. Each dataset contains all epochs from the minority class and a disjoint set of epochs from the majority class. Each of these individual datasets can now be used to construct a separate model.

E. Long Short-Term Memory Neural Networks

Each dataset resulting from the balanced bootstrapping procedure is modeled using a powerful model known as a Long Short-Term Memory (LSTM) [38] neural network. It is used to capture temporal information and accurately model the data. LSTM networks are a type of RNN based on LSTM cells.

The network architecture for a single instance of the LSTM is shown in Fig. 4. This architecture is similar to other architectures used for sequence modeling [31], [39]. It consists of an LSTM layer with n_1 cells followed by a dropout layer with dropout probability p_1. Dropout layers are used to improve the generalization of the network towards unseen data [40]. Finally,

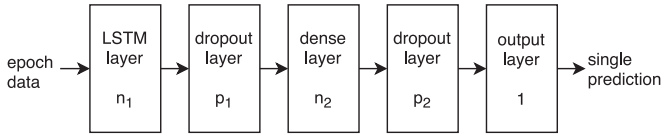


Fig. 4. Architecture of a single instance of the proposed sleep apnea detection model using LSTM cells.

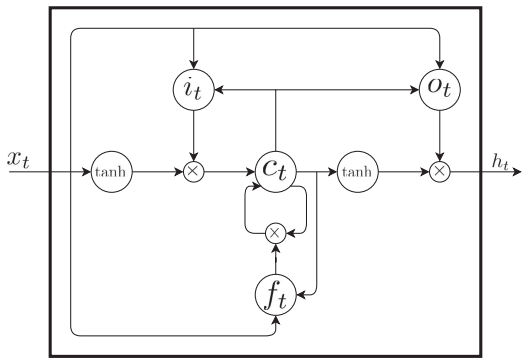


Fig. 5. Flowchart of the LSTM cells used in this work.

a dense layer with n_2 cells is appended followed by a dropout layer with dropout probability p_2 and the output prediction cell with a sigmoid activation function. This activation function results in an output that can be interpreted as the probability that the input epoch contains apnea. This architecture is replicated for each LSTM model in Fig. 1.

A flowchart of the cell used in this work is shown in Fig. 5 and the corresponding equations are shown in (1).

$$\begin{aligned} i_t &= \sigma(x_t W_{x_i} + h_{t-1} W_{h_i} + c_{t-1} W_{c_i} + b_i) \\ f_t &= \sigma(x_t W_{x_f} + h_{t-1} W_{h_f} + c_{t-1} W_{c_f} + b_f) \\ c_t &= f_t c_{t-1} + i_t \tanh(x_t W_{x_c} + h_{t-1} W_{h_c} + b_c) \\ o_t &= \sigma(x + tW_{x_o} + h_{t-1} W_{h_o} + c_t W_{c_o} + b_o) \\ h_t &= o_t \tanh(c_t) \end{aligned} \tag{1}$$

In these equations, σ is the logistic sigmoid function, x_t represents the input sequence x at time t consisting of respiratory data measured at 5 Hz and h_t represents the hidden state at time t. The input gate, forget gate and output gate are represented by i, f, o respectively. The cell and cell input activation vectors are represented by o, c. The weight matrices are represented by W and the bias terms by b.

To tune the hyperparameters n_1, n_2, p_1 and p_2 of the network, Bayesian optimization (BO) is used which is a powerful strategy to optimize hyperparameters of medical machine learning models [33], [41], [42]. It converges the network architecture to an optimal design for accurate prediction of unseen data. In this work, the Efficient Global Optimization algorithm is used with the Expected Improvement acquisition function [43]. More details about BO of hyperparameters are given in [41].

For each of the datasets generated by the balanced bootstrapping procedure, a network as shown in Fig. 4 is trained using minibatches of 32 epochs, consisting of 16 positive and 16

negative epochs. The weights W and b of the network are optimized using the adadelta optimizer [44] and the loss function is the binary crossentropy as defined by:

$$\text{loss} = \sum_{i=1}^{N} -(y_i \log(p_i) + (1 - y_i) \log(1 - p_i))$$

where N represents the number of samples to compute the metric on, y_i indicates the true binary label of sample i and p_i represents the predicted probability for sample i.

F. Aggregated Prediction

To test the presence of a sleep apnea event in a new epoch of data, it is pre-processed as in Section III-B and passed through each of the trained LSTM models in Fig. 1. Each LSTM model outputs the probability of an apnea event because of the sigmoid activation function. The resulting probability estimates of each separate LSTM model are then aggregated into a single probability prediction per epoch by averaging. An epoch can be labeled as either apnea or non-apnea by determining if the estimated averaged probability $p \geq 0.5$. This classifier can be further fine-tuned for specific use-cases by defining a threshold $p \geq \tau$ and optimizing the value of τ.

To get a measure of the severity of apnea of a patient, the Apnea Hypopnea Index (AHI) is computed as the number of apnea events longer than 10 seconds divided by the total sleep time following the official AASM guidelines [6]. A prediction of the probability of sleep apnea for each second of the signal is made using the aggregated probability of sleep apnea from the LSTM models. These generated annotations are then used to compute the AHI. Using this score, patients are classified with normal breathing, mild, moderate or severe sleep apnea:

- Normal breathing: AHI ≤ 5
- Mild sleep apnea: $5 \leq$ AHI ≤ 15
- Moderate sleep apnea: $15 \leq$ AHI ≤ 30
- Severe sleep apnea: AHI ≥ 30

IV. Experimental Setup and Method

A. Dataset

To validate the proposed method, the SHHS-1 dataset [11] is used, which contains data of 5804 adults of age 40 and older. The comprehensive size of this dataset makes it possible to test in a reliable way whether the algorithms are able to generalize to many different patients. Out of these 5804 patients, 2100 patients were sequentially selected. The only selection requirement was having at least six hours of useful data. This set is then split up five times in disjoint training sets of 100 patients and test sets of 2000 patients. A training set of 100 patients provides enough variation in the patients while keeping the computational burden for the model low.

The dataset consists of 1008 female and 1092 male patients with mean age 62.5 ± 12.6 (standard deviation) years, mean weight 74.0 ± 19.3 kg and mean BMI $27.2 \pm 5.3\,\text{kg/m}^2$. The mean recording length is 10.1 ± 1.6 hours with a mean sleep time of 6.2 ± 1.0 hours. There are 35 patients with normal breathing (mean AHI $= 3.8 \pm 1.1$), 450 patients with mild sleep apnea (mean AHI $= 10.7 \pm 2.7$), 815 patients with moderate sleep apnea (mean AHI $= 22.1 \pm 4.3$) and 800 patients with severe apnea (mean AHI $= 44.1 \pm 12.2$).

The SHHS-1 dataset contains a variety of physiological signals measured for each patient including respiratory, cardiovascular and oxygen saturation signals. In this work, the focus is on sleep apnea detection in respiratory signals. A recent study on the comparison of different respiratory signals showed that direct measurements from respiratory belts around the abdomen and thorax resulted in the best predictive performance [33]. Hence, the proposed algorithm is tested using these two respiratory signals. To also test the performance of the algorithm on indirect measurements, the EDR signal is includes as well. Each of these signals are tested in a separately trained model:

- *Abdores*: Abdominal respiratory belt below the lower edge of the left ribcage.
- *Thorres*: Thoracic respiration belt below left armpit.
- *EDR*: ECG derived respiration signal by filtering the ECG signal with a cut-off frequency of 0.4 Hz and high-pass filtering this signal with a cut-off frequency of 0.2 Hz [45].

The annotations of sleep apnea in the SHHS-1 dataset are based on the SHHS method [11].

B. Benchmark Methods

To compare the performance of the proposed method versus the state-of-the-art, three model types are included in the experimental setup:

- *Standard machine learning:* An Artificial Neural Network (ANN) model [26], Logistic Regression (LR) model [23] and Random Forest (RF) model [46] are evaluated and compared as these are frequently used in sleep apnea or other medical use-cases.
- *Temporal machine learning:* An LSTM network, similar to the one introduced in Fig. 4, but now the inputs are human-engineered features instead of raw respiratory signals. It is denoted as F-LSTM.
- *New method:* The proposed new method of Fig. 1, denoted as LSTM, which uses raw respiratory signals that have only been noise filtered.

The hyperparameters of all models are tuned using BO. The implementation of the BO algorithm is based on the GPyOpt Python library [47]. To train and test the models with the respiratory signals, the same pre-processing steps as Fig. 1 are performed and typical discriminative features for sleep apnea, sleep studies and biomedical health in general, are extracted, both in the time-domain as well as the frequency-domain [24], [27], [33]. An overview of features is provided in Table I.

For the standard machine learning models (ANN, LR and RF), the set of 100 training patients is used in a 5-fold cross-validation during optimization of the models to prevent overfitting. For the temporal machine learning models (F-LSTM and LSTM), the set of 100 training patients is split up per-patient in a training, test and validation set for training and optimizing the model, as is typically done in deep learning to avoid large computational demands. These methods of training and optimization adhere

TABLE I
OVERVIEW OF HUMAN-ENGINEERED FEATURES USED IN BENCHMARK METHODS

Origin	Feature
Time-domain	mean, standard deviation, skewness, area under absolute value
Respiratory peaks	mean of heights, standard deviation of heights, skewness of heights, number of peaks, mean inter-peak distance, standard deviation of peak-distance, skewness of inter-peak distance, sum of peak heights
Frequency-domain	peak frequency, mean frequency, central frequency, band power

to the recommendations and common practices for each model type.

As this study aims to analyze the predictive power of several algorithms for the automated detection of sleep apnea in clinical settings, no epochs are removed from the training set, nor from the test set.

C. Evaluation Criteria

Sleep apnea detection algorithms are evaluated using a variety of metrics. Typically, the per-epoch classification accuracy is calculated using metrics such as the sensitivity, also known as recall, and specificity. These are based on the number of True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN). Another commonly used metric is the classification accuracy.

$$\text{sensitivity} = \text{TP}/(\text{TP} + \text{FN})$$
$$\text{specificity} = \text{TN}/(\text{TN} + \text{FP})$$
$$\text{accuracy} = (\text{TP} + \text{TN})/(\text{FP} + \text{TN} + \text{TP} + \text{FN})$$

Preferably, all these metrics have a high score. However, they can easily be influenced by changing the decision threshold τ in Section III-F. Therefore, a more complete assessment is achieved by computing the Receiver Operator Characteristics (ROC) and the associated area under the curve (AUROC) [23] as these metrics summarize the results for all thresholds τ. The ROC curve is created by determining the unique pairs of sensitivity and specificity for all possible thresholds τ and plotting this in a graph of sensitivity versus $1 -$ specificity [48]. In order to compute this, the models need to output a probability p of an event. All models used in this experimental setup are configured to output this probability. The area under the curve can be computed by integrating across all thresholds.

Many works in literature report very high sensitivity and specificity scores. However, when dealing with datasets that have an imbalance in the number of positive vs negative samples, such metrics can provide misleading insights [35], [36]. As this imbalance is certainly the case in sleep apnea, it is advisable to further take into account the precision and negative-predictive-value (NPV) metrics.

$$\text{precision} = \text{TP}/(\text{TP} + \text{FP})$$
$$\text{NPV} = \text{TN}/(\text{TN} + \text{FN})$$

TABLE II
OPTIMAL HYPERPARAMETERS FOR LSTM MODEL RESULTING FROM THE BAYESIAN OPTIMIZATION PROCEDURE

signal	n1	n2	p1	p2
abdores	100	50	0.5	0.5
thorres	100	50	0.5	0.5
EDR	50	20	0.14	0.27

Given that the precision is also susceptible to an arbitrary decision threshold, the possible combinations of precision and sensitivity can be summarized in the precision-recall curve and the area under this curve (AUPRC). The PR-curve is computed using a similar method as the ROC curve but instead of pairs of sensitivity and specificity, it uses pairs of precision and recall.

Next to being able to accurately predict the per-epoch annotation label, it is also important to be able to predict the per-patient AHI. Hence, the classification accuracy per AHI class is also computed. This reflects the accuracy of categorizing a patient in any of the 4 AHI classes. To compute these scores, the decision threshold τ, discussed in Section III-F, is optimized. The threshold τ determines when a specific probability of apnea is sufficient to flag the epoch as containing an apnea event. The AHI classification of the 100 training patients is computed for various decision thresholds τ between 0 and 1. The threshold τ leading to the best classification accuracy for the training patients is used to compute the AHI for the test patients. This is repeated across the five iterations of the experiment.

V. RESULTS AND DISCUSSION

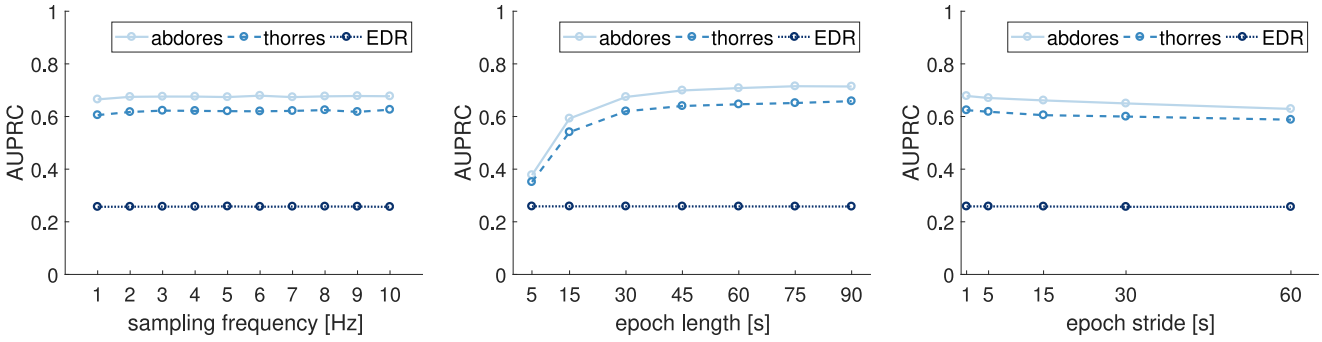
In the following discussion, the parameters for the proposed LSTM model are presented and discussed. Next, the per-epoch metrics are discussed and compared to literature. The per-subject metrics are also discussed and the advantages of the balanced bootstrapping procedure are analyzed. Finally, the overall performance of the model is evaluated.

A. Parameters of the Model

The optimal hyperparameters of the LSTM model for each of the respiration signals, generated by the BO procedure, are shown in Table II. In addition to the parameters that were optimized, other parameters were considered fixed, namely: the sampling frequency f_s of the input data, the epoch length l_{epoch} and the stride s_{epoch} between consecutive epochs. The AUPRC on the validation dataset is computed for a range of the parameters. The effect of varying these parameters is shown in Fig. 6.

Changing the sampling frequency f_s has little to no effect on the performance of the model as demonstrated by Fig. 6a. This is because all signals have been low-pass filtered with a cutoff frequency of 0.7 Hz during the pre-processing step, as discussed in Section III-B.

On the other hand, changing the length of the epochs does have a considerable impact on the performance of the model as demonstrated by Fig. 6b. For the abdores and thorres signals, an improvement can be seen up to a length of 30 seconds. After that, the performance gain is minimal although the



(a) Sampling frequency (b) Epoch length (c) Epoch stride

Fig. 6. Influence of the model parameter settings on the AUPRC of the validation dataset.

computational requirements drastically increase. This is mainly due to the LSTM model being used. Although the longer epochs contain more valuable data, the results show that the LSTM model can no longer successfully exploit this extra information. In addition, increasing the epoch length leads to a longer training process.

Lastly, when increasing the stride between epochs, there is a slight drop in performance as demonstrated by Fig. 6c. When the stride is small more data is available to train the model on. When choosing a larger stride, the amount of epochs in the training data drops.

B. Per-Epoch Score Evaluation

Table III provides an overview of all per-epoch evaluation metrics for each model and each respiration signal separately. The values represent the mean and standard deviation across the five experiment iterations. The per-epoch results show that in general, the LSTM model outperforms the state-of-the-art models, especially when evaluating the AUPRC. This metric demonstrates a considerable reduction in false positives for the proposed model in comparison to the benchmark methods. However, the amount of false positives is still quite high across all tested methods and this can be attributed to several causes:

- Typically, only episodes longer than 10 seconds are annotated. However, the models make decisions on a per-second basis and hence also detect shorter respiratory disturbances. Luckily, these short false positives can easily be filtered out by a post-processing step.
- As mentioned in Section IV-B, no manual cleaning has been performed on the data to accurately reflect a real-life measurement of a patient. Only noise-canceling methods have been applied, but these cannot fully remove all measurement noise. The models can interpret this noise as irregular breathing and flag the epoch as an apnea episode. The proposed LSTM model is more robust against noise as it evaluates the full data of the epoch to make a decision instead of only using a set of summarized features. Manual epoch removal can significantly improve these results as shown in other works, but is not representative of the use-case this work aims to evaluate.
- All models aim to detect sleep apnea events based on respiratory signals. However, trained sleep technicians also take into account other signals such as the oximetry data. The proposed method is a way of quick screening to assist trained staff.
- The human annotations of the position of the sleep apnea events are not exact to the second. However, the model evaluation requires the position to be as accurate as the annotations. Often, the models start detecting several seconds too early or too late. When computing the AHI, used in analysis of results per patient, this does not influence the results in any way.

These results can be compared with scores reported in literature. Table IV provides an overview of several other studies found in literature for automated prediction of sleep apnea using respiratory signals or the ECG signal. The standard machine learning models included in the comparison are the Support Vector Machine (SVM) model [49], the LR model [50] and the RF model [51]. The temporal machine learning models include the LSTM model [29], [30] and the SVM-based Hidden Markov Model (HMM) [27]. The comparison also includes a rule-based model [52] and a deep-learning Convolutional Neural Network (CNN) model [53]. For an excellent overview of other models in sleep apnea detection, we refer to two recent review papers [12], [54].

The comparison shows how the other studies typically achieve high scores for apnea classification using the ECG signal when the model is based on human-engineered features. This is in contrast to scores for the EDR method used in this work. A major limitation of the ECG signal however, is the influence of other illnesses on the sleep apnea analyses. The EDR signal is less susceptible to this but is still influenced by noise. When comparing the studies that use respiration signals, the results are comparable to our proposed LSTM method. Note however, that it is difficult to fully compare these results as they are computed on different datasets and with different experimental setups. Furthermore, without the AUPRC metric, it is difficult to analyze the performance of the algorithm with regards to false positive predictions. As there is a large data imbalance, algorithms with more false positives than true positives can still achieve good scores using the other metrics. With this AUPRC

TABLE III

EVALUATION METRICS FOR THE DIFFERENT SLEEP APNEA DETECTION MODELS FOR EACH RESPIRATION SIGNAL, AGGREGATED ACROSS ALL EPOCHS OF THE 2000 UNSEEN TEST PATIENTS AND AVERAGED ACROSS THE FIVE EXPERIMENT ITERATIONS RESULTING IN THE MEAN AND STANDARD DEVIATION ESTIMATES. ALL METRICS ARE EXPRESSED IN PERCENTAGES. IN GENERAL, THE PROPOSED LSTM MODEL OFFERS THE BEST RESULTS

(a) Non-temporal models

	Abdores			**Thorres**			**EDR**		
%	***ANN***	***LR***	***RF***	***ANN***	***LR***	***RF***	***ANN***	***LR***	***RF***
sensitivity	66.0 ± 2.7	31.4 ± 8.7	17.5 ± 1.1	70.0 ± 3.4	76.6 ± 3.3	17.2 ± 1.8	40.2 ± 24.8	96.6 ± 4.0	7.7 ± 1.4
specificity	55.9 ± 2.4	79.5 ± 6.7	95.6 ± 0.8	56.4 ± 4.2	40.9 ± 4.9	96.3 ± 0.7	63.9 ± 23.3	5.4 ± 6.1	99.2 ± 0.4
precision	23.8 ± 0.6	24.5 ± 1.5	45.6 ± 0.0	25.1 ± 1.0	21.2 ± 0.7	49.2 ± 2.8	18.9 ± 0.9	17.6 ± 0.4	67.7 ± 8.2
NPV	88.7 ± 0.5	84.8 ± 0.6	84.8 ± 0.2	90.0 ± 0.5	89.2 ± 0.2	84.8 ± 0.3	83.9 ± 0.9	88.4 ± 0.6	83.7 ± 0.2
accuracy	57.7 ± 1.6	71.2 ± 4.1	82.1 ± 0.6	58.7 ± 3.0	47.0 ± 3.5	82.6 ± 0.4	59.8 ± 14.2	21.2 ± 4.4	83.4 ± 0.3
AUPRC	20.8 ± 0.1	19.0 ± 0.5	23.3 ± 0.6	22.2 ± 0.4	20.1 ± 0.3	22.7 ± 0.6	18.0 ± 0.5	17.5 ± 0.5	20.8 ± 0.6
AUROC	58.7 ± 0.6	53.0 ± 1.0	54.8 ± 0.4	62.1 ± 0.7	57.7 ± 0.5	55.4 ± 0.6	51.6 ± 1.3	50.8 ± 0.8	52.4 ± 0.4

(b) Temporal models

	abdores		**thorres**		**EDR**	
%	***fLSTM***	***LSTM***	***fLSTM***	***LSTM***	***fLSTM***	***LSTM***
sensitivity	57.9 ± 8.6	62.3 ± 2.9	62.9 ± 3.5	67.8 ± 2.5	48.8 ± 10.2	52.1 ± 0.0
specificity	73.9 ± 10.0	80.3 ± 2.3	77.2 ± 4.5	76.5 ± 2.3	60.8 ± 12.5	61.8 ± 1.4
precision	33.0 ± 5.8	39.9 ± 1.9	36.8 ± 3.1	37.7 ± 1.6	21.1 ± 2.2	22.1 ± 0.2
NPV	89.5 ± 0.9	91.1 ± 0.4	90.9 ± 0.4	91.9 ± 0.4	85.0 ± 0.6	86.1 ± 0.2
accuracy	71.1 ± 6.8	77.2 ± 1.4	74.7 ± 3.1	75.0 ± 1.4	58.7 ± 8.6	60.1 ± 0.9
AUPRC	36.4 ± 2.4	45.3 ± 1.2	43.9 ± 0.2	48.0 ± 1.0	22.1 ± 0.9	22.7 ± 0.2
AUROC	71.5 ± 1.7	77.5 ± 0.5	76.9 ± 0.8	79.7 ± 0.4	57.6 ± 1.7	58.8 ± 0.2

TABLE IV

COMPARISON METRICS FROM SEVERAL OTHER SLEEP APNEA STUDIES USING EITHER ECG OR RESPIRATORY DATA. EVENT TYPES ARE CLASSIFIED AS APNEA (A), OBSTRUCTIVE APNEA (O), HYPOPNEA (H) OR NO APNEA (N). NO STUDIES REPORT THE VALUABLE AUPRC METRIC. ALL VALUES ARE REPRESENTED AS PERCENTAGES

study	**model**	**signal**	**event**	**dataset**	**granularity**	**sensitivity**	**specificity**	**precision**	**npv**	**accuracy**	**AUPRC**	**AUROC**
[29]	LSTM	ECG	AH/N	35+45	epoch	99.9	100.0	-	-	99.9	-	-
[30]	LSTM	ECG	O/N	35	recording	-	-	-	-	97.8	-	-
[27]	SVM-HMM	ECG	O/N	70	subject	82.6	88.4		-	86.2	-	94.0
[49]	SVM	resp.	AH/N	4	epoch	93.2	88.9	90.0	-	89.9	-	-
[50]	LR	resp.	AH/N	148	subject	88.0	70.8	-	-	82.4	-	90.3
[51]	RF	resp.	A/N	8	epoch	-	-	-	-	92.8	-	-
[53]	CNN	resp.	O/N	100.0	epoch	74.7	-	74.5	-	74.7	-	-
[52]	rule-based	resp.	A/N	100.0	epoch	83.6	72.3	-	-	-	-	-

metric, a more complete assessment of the performance can be made.

C. Per-Patient Score Evaluation

When physicians analyze a patient to estimate the severity of sleep apnea, the individual per-epoch scores are not used. Instead, they base their decisions on the aggregated AHI metric. Fig. 7 shows the confusion matrices for the AHI classification for each benchmark method computed using the abdominal respiratory signal. The confusion matrices represent the mean (and standard deviation) across the five experiment iterations. The figure demonstrates that the predictions of the LSTM model are more concentrated around the actual target class than for the other models. Importantly, no severe apnea cases where classified as normal breathing with the LSTM model. The predictions for the other models are more biased.

The results of the confusion matrices can further be summarized in a classification accuracy graph as shown in Fig. 8 for all respiration signals and all models. The graph shows that in general, all models perform much better with the moderate and severe apnea cases than the normal or mild cases when using respiratory data. This can be traced back to different dynamics in respiration for patients with normal breathing when compared to patients with severe apnea. Studies have shown that patients with severe obstructive apnea have a higher activity in the sympathetic nervous system [55], [56]. When the activity in the sympathetic nervous system increases, the respiration rate also increases. As the amount of normal (1.65%) or mild (22.70%) apnea patients is much more limited in the database than the amount of moderate (47.70%) or severe (34.95%) patients, the normal respiration patterns are underrepresented in the dataset and the model is unable to sufficiently learn these dynamics in comparison to the dynamics of severe apnea patients.

When analyzing the AHI classification accuracy, the performance of the other benchmark models is improved when compared to only analyzing the per-epoch metrics. This is because the per-epoch metrics require predictions to be accurate for the exact location/duration of the event.

Fig. 8c shows the classification accuracy for the EDR signal. These results show that the accuracy of the model with derived respiration is less then when using direct respiration.

D. Effect of Balanced Bootstrapping

The confusion matrices can also be used to assess the variation of predictions across the different bootstraps from Section III-D. Fig. 9 demonstrates the mean and standard

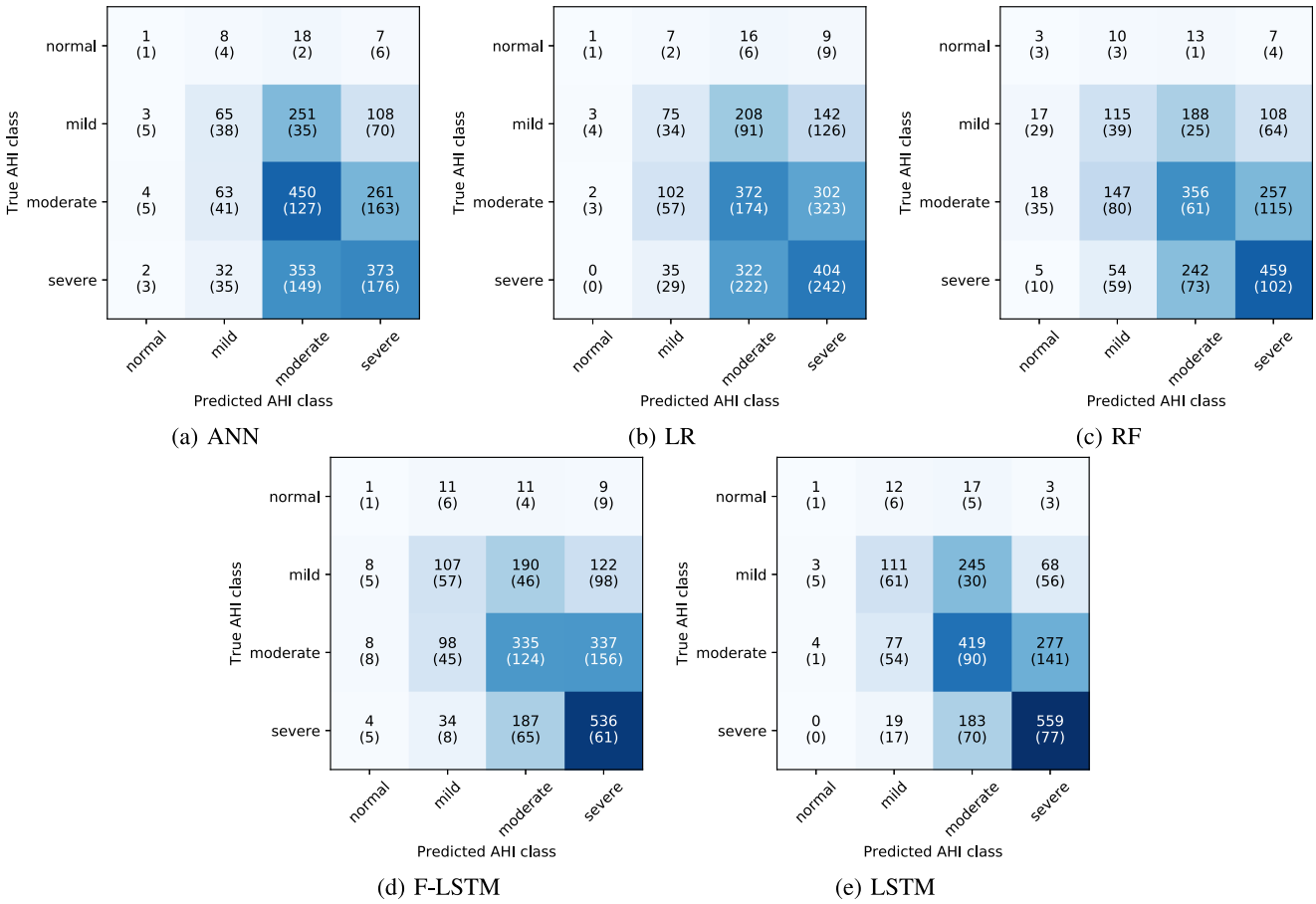


Fig. 7. Confusion matrices for the prediction of AHI class, using the *abdores* respiratory signal with the different machine learning models. The values represent the mean and standard deviation across the five experiment iterations. The results of the LSTM model classifications are most concentrated around their actual target classes.

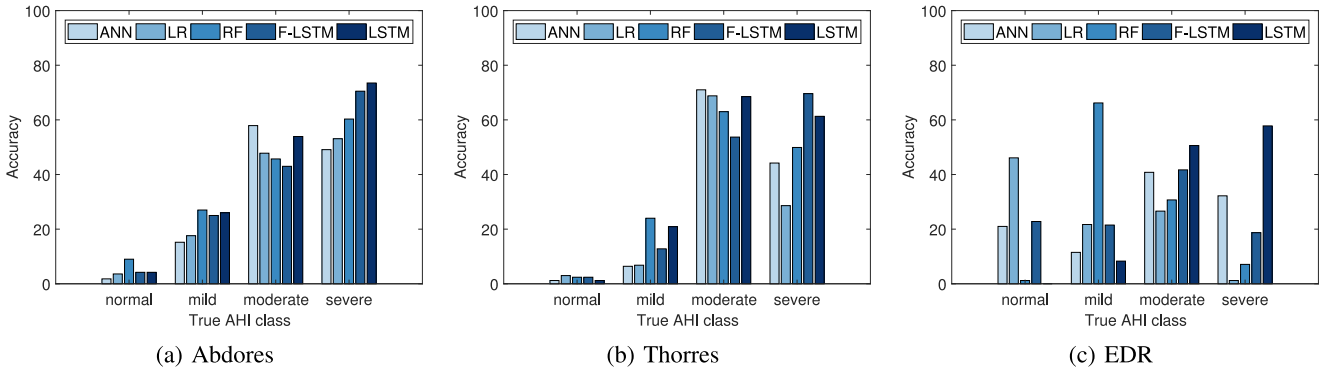
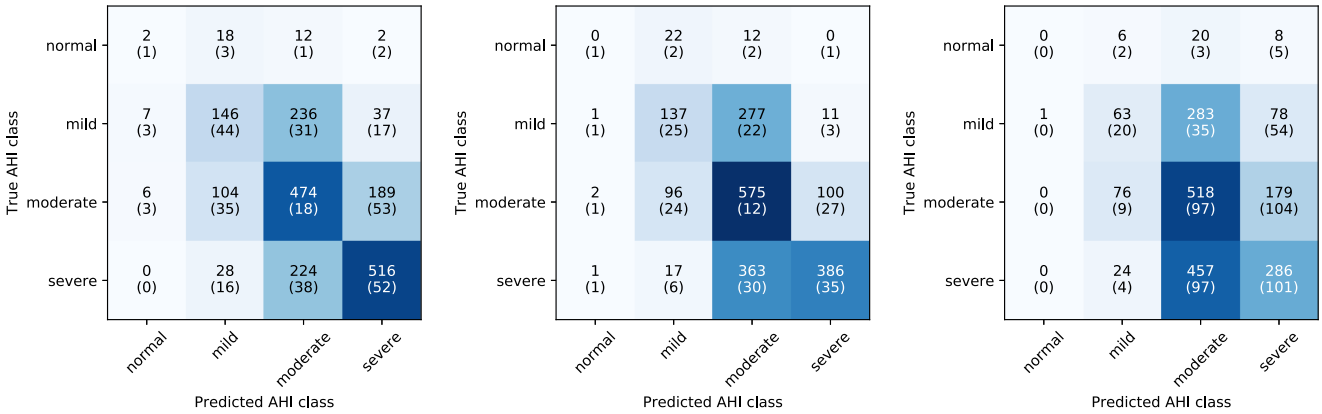


Fig. 8. Classification accuracy for each AHI class and each respiratory signal averaged across the five experiment iterations. The LSTM model offers the best overall performance. There is a considerable drop in performance for the EDR signal.

deviation of AHI classification in confusion matrices, computed using the three different respiration signals with the LSTM model across the different balanced bootstrapping predictions in one of the five experiment iterations. The figures demonstrate an increased variation in predictive performance across the bootstrapped models for the EDR signal when compared to the models with respiratory signals, indicating an increased advantage of the balanced bootstrapping procedure when the signal is less clean.

E. Overall Performance

When comparing the metrics and analyses, the LSTM model, trained on respiratory signals without any human feature



(a) Abdores (b) Thorres (c) EDR

Fig. 9. Confusion matrices indicating the mean and standard deviation of AHI classification across the different balanced boostrap predictions. There is a larger variation for the EDR signal than for the respiratory signals.

engineering outperforms the current state-of-the-art methods for sleep apnea detection in respiratory signals. When comparing the specialized metrics for imbalanced data, the LSTM model shows a significant increase over the state-of-the-art models. Because the LSTM model learns and predicts from raw data with only minimal noise filtering, it is able to provide more stable predictions in practical settings than the other models, resulting in less false positives.

The SHHS-1 dataset was used in this experiment for its large size, enabling us to test in a reliable way if the developed methods are able to generalize to new, unseen patients with a variety of characteristics. However, since the development of this dataset, there has been a considerable improvement in apnea annotation criteria for the reduction of inter-rater and intra-rater variability as demonstrated by recent studies [10]. As annotations using these new criteria are more consistent, the predictions of models trained using these new criteria will also be more consistent, which will result in higher performance metrics.

The proposed model offers a powerful method for quickly analyzing the recording based on respiratory information alone. It represents a first important step towards a fully automated sleep apnea detection method. Furthermore, it also provides a method of quickly indicating interesting epochs for trained staff, allowing them to focus on the interesting sections during the night, and allowing more patients to be evaluated.

VI. Conclusion and Future Work

As sleep apnea is one of the most common sleep disorders and the consequences can be very severe, more patients need to be analyzed and automatic detection methods are needed. Many such methods have been proposed over the years. These typically use human-engineered features. In this work, a novel method of training LSTM networks on the respiratory signal itself, i.e., without the need for manual feature engineering, is proposed. The method is able to detect OSA, CSA as well as hypopnea. Preprocessing the signals to extract respiratory information combined with efficient usage of the data via the balanced bootstrapping scheme enables the training of LSTM networks on long sequences of respiratory signals, which results in a more robust and more accurate model when analyzing new patients.

The analysis is performed using typical sleep apnea metrics as well as specialized metrics for imbalanced data. The results of evaluating this model on five sets of 2000 unseen patients show a considerable improvement when compared to the current state-of-the-art. There is a significant increase in per-epoch performance as well as in accuracy for AHI-based classification. This study also demonstrates the importance of using specialized metrics for imbalanced data when assessing the performance of machine learning models for the detection of sleep apnea. These results provide valuable insights for the further development of automated sleep apnea screening tools.

When analyzing sleep quality, other events such as teeth grinding and snoring are also equally important. In future work, our model will be extended to also include these other types of event to provide a complete and accurate sleep analysis method.

References

[1] C. Guilleminault, A. Tilkian, and W. C. Dement, "The sleep apnea syndromes," *Annu. Rev. Med.*, vol. 27, no. 1, pp. 465–484, 1976.
[2] R. Heinzer *et al.*, "Prevalence of sleep-disordered breathing in the general population: The hypnolaus study," *Lancet Respiratory Med.*, vol. 3, no. 4, pp. 310–318, 2015.
[3] V. K. Somers *et al.*, "Sleep apnea and cardiovascular disease," *Circulation*, vol. 118, no. 10, pp. 1080–1111, 2008.
[4] H. K. Yaggi, J. Concato, W. N. Kernan, J. H. Lichtman, L. M. Brass, and V. Mohsenin, "Obstructive sleep apnea as a risk factor for stroke and death," *New England J. Med.*, vol. 353, no. 19, pp. 2034–2041, 2005.
[5] A. Sassani, L. J. Findley, M. Kryger, E. Goldlust, C. George, and T. M. Davidson, "Reducing motor-vehicle collisions, costs, and fatalities by treating obstructive sleep apnea syndrome," *Sleep*, vol. 27, no. 3, pp. 453–458, 2004.
[6] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. L. Marcus, and B. V. Vaughn, "The AASM manual for the scoring of sleep and associated events," in *Rules, Terminology and Technical Specifications*. Darien, IL, USA: American Academy of Sleep Medicine, 2012.
[7] W. W. Flemons, N. J. Douglas, S. T. Kuna, D. O. Rodenstein, and J. Wheatley, "Access to diagnosis and treatment of patients with suspected sleep apnea," *Amer. J. Respiratory Crit. Care Med.*, vol. 169, no. 6, pp. 668–672, 2004.

[8] D. Bliwise, N. G. Bliwise, H. C. Kraemer, and W. Dement, "Measurement error in visually scored electrophysiological data: Respiration during sleep," *J. Neuroscience Methods*, vol. 12, no. 1, pp. 49–56, 1984.

[9] N. A. Collop, "Scoring variability between polysomnography technologists in different sleep laboratories," *Sleep Med.*, vol. 3, no. 1, pp. 43–47, 2002.

[10] S. T. Kuna *et al.*, "Agreement in computer-assisted manual scoring of polysomnograms across sleep centers," *Sleep*, vol. 36, no. 4, pp. 583–589, 2013.

[11] S. F. Quan *et al.*, "The sleep heart health study: Design, rationale, and methods," *Sleep*, vol. 20, no. 12, pp. 1077–1085, 1997.

[12] M. Uddin, C. Chow, and S. Su, "Classification methods to detect sleep apnea in adults based on respiratory and oximetry signals: A systematic review," *Physiological Meas.*, vol. 39, no. 3, 2018, Art. no. 03TR01.

[13] N. A. Collop *et al.*, "Clinical guidelines for the use of unattended portable monitors in the diagnosis of obstructive sleep apnea in adult patients." *J. Clin. Sleep Med.*, vol. 3, no. 7, pp. 737–747, 2007.

[14] G. C. Gutierrez-Tobal, D. Alvarez, A. Crespo, F. Del Campo, and R. Hornero, "Evaluation of machine-learning approaches to estimate sleep apnea severity from at-home oximetry recordings," *IEEE J. Biomed. Health Inform.*, to be published.

[15] N. Netzer, A. H. Eliasson, C. Netzer, and D. A. Kristo, "Overnight pulse oximetry for sleep-disordered breathing in adults: A review," *Chest*, vol. 120, no. 2, pp. 625–633, 2001.

[16] M. Bsoul, H. Minn, and L. Tamil, "Apnea medassist: Real-time sleep apnea monitor using single-lead ECG," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 3, pp. 416–427, May 2011.

[17] G. B. Moody *et al.*, "Clinical validation of the ECG-derived respiration (EDR) technique," *Comput. Cardiology*, vol. 13, no. 3, pp. 507–510, 1986.

[18] W. Zhao, H. Ni, X. Zhou, Y. Song, and T. Wang, "Identifying sleep apnea syndrome using heart rate and breathing effort variation analysis based on ballistocardiography," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2015, pp. 4536–4539.

[19] I. Sadek, E. Seet, J. Biswas, B. Abdulrazak, and M. Mokhtari, "Nonintrusive vital signs monitoring for sleep apnea patients: A preliminary study," *IEEE Access*, vol. 6, pp. 2506–2514, 2018.

[20] 2018. [Online]. Available: https://physionet.org/challenge/2018/

[21] G. Sannino, I. De Falco, and G. De Pietro, "An automatic rules extraction approach to support OSA events detection in an mhealth system," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 5, pp. 1518–1524, Sep. 2014.

[22] B. Yılmaz, M. H. Asyalı, E. Arıkan, S. Yetkin, and F. Özgen, "Sleep stage and obstructive apneaic epoch classification using single-lead ECG," *Biomed. Eng. Online*, vol. 9, no. 1, pp. 1–14, 2010.

[23] J. V. Marcos, R. Hornero, D. Álvarez, F. del Campo, and C. Zamarrón, "Assessment of four statistical pattern recognition techniques to assist in obstructive sleep apnoea diagnosis from nocturnal oximetry," *Med. Eng. Phys.*, vol. 31, no. 8, pp. 971–978, 2009.

[24] P. De Chazal, C. Heneghan, E. Sheridan, R. Reilly, P. Nolan, and M. O'Malley, "Automatic classification of sleep apnea epochs using the electrocardiogram," in *Proc. IEEE Comput. Cardiology*, 2000, pp. 745–748.

[25] S. Gutta, Q. Cheng, H. Nguyen, and B. Benjamin, "Cardiorespiratory model-based data-driven approach for sleep apnea detection," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 4, pp. 1036–1045, Jul. 2018.

[26] P. Várady, T. Micsik, S. Benedek, and Z. Benyó, "A novel method for the detection of apnea and hypopnea events in respiration signals," *IEEE Trans. Biomed. Eng.*, vol. 49, no. 9, pp. 936–942, Sep. 2002.

[27] C. Song, K. Liu, X. Zhang, L. Chen, and X. Xian, "An obstructive sleep apnea detection approach using a discriminative hidden Markov model from ECG signals," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1532–1542, Jul. 2016.

[28] D. Novák, K. Mucha, and T. Al-Ani, "Long short-term memory for apnea detection based on heart rate variability," in *Proc. 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2008, pp. 5234–5237.

[29] R. K. Pathinarupothi, R. Vinaykumar, E. Rangan, E. Gopalakrishnan, and K. Soman, "Instantaneous heart rate as a robust feature for sleep apnea severity detection using deep learning," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Inform.*, 2017, pp. 293–296.

[30] M. Cheng, W. J. Sori, F. Jiang, A. Khan, and S. Liu, "Recurrent neural network based classification of ECG signal features for obstruction of sleep apnea detection," in *Proc. IEEE Int. Conf. Comput. Sci. Eng. Embedded Ubiquitous Comput.*, 2017, vol. 2, pp. 199–202.

[31] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to diagnose with LSTM recurrent neural networks," 2015, arXiv:1511.03677.

[32] T. Van Steenkiste *et al.*, "Accurate prediction of blood culture outcome in the intensive care unit using long short-term memory neural networks," *Artif. Intell. Med.*, p. 6, 2019.

[33] T. Van Steenkiste *et al.*, "Systematic comparison of respiratory signals for the automated detection of sleep apnea," in *Proc. EMBS 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2018, pp. 449–452.

[34] D. A. Hettrick and T. M. Zielinski, "Bioimpedance in cardiovascular medicine," in *Encyclopedia of Medical Devices and Instrumentation*. Hoboken, NJ, USA: Wiley, 2006.

[35] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[36] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review," *GESTS Int. Trans. Comput. Sci. Eng.*, vol. 30, no. 1, pp. 25–36, 2006.

[37] B. C. Wallace, K. Small, C. E. Brodley, and T. A. Trikalinos, "Class imbalance, redux," in *Proc. IEEE 11th Int. Conf. Data Mining*, 2011, pp. 754–763.

[38] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[39] A. M. Dai and Q. V. Le, "Semi-supervised sequence learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 3079–3087.

[40] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[41] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2951–2959.

[42] T. Van Steenkiste *et al.*, "Automated assessment of bone age using deep learning and Gaussian process regression," in *Proc. EMBS 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2018, pp. 674–677.

[43] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient global optimization of expensive black-box functions," *J. Global Optim.*, vol. 13, no. 4, pp. 455–492, 1998.

[44] M. D. Zeiler, "Adadelta: An adaptive learning rate method," 2012, arXiv:1212.5701.

[45] G. D. Clifford, *Advanced Methods and Tools for ECG Data Analysis*. Norwood, MA, USA: Artech House, 2009.

[46] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[47] The GPyOpt authors, "GPyOpt: A Bayesian optimization framework in python," 2016. [Online]. Available: http://github.com/SheffieldML/GPyOpt

[48] A. K. Akobeng, "Understanding diagnostic tests 3: Receiver operating characteristic curves," *Acta Paediatrica*, vol. 96, no. 5, pp. 644–647, 2007.

[49] B. L. Koley and D. Dey, "Real-time adaptive apnea and hypopnea event detection methodology for portable sleep apnea monitoring devices," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 12, pp. 3354–3363, Dec. 2013.

[50] G. C. Gutiérrez-Tobal, R. Hornero, D. Álvarez, J. V. Marcos, and F. del Campo, "Linear and nonlinear analysis of airflow recordings to help in sleep apnoea–hypopnoea syndrome diagnosis," *Physiological Meas.*, vol. 33, no. 7, pp. 1261–1275, 2012.

[51] C. Avcı and A. Akbaş, "Sleep apnea classification based on respiration signals by using ensemble methods," *Bio-medical Mater. Eng.*, vol. 26, no. s1, pp. S1703–S1710, 2015.

[52] N. Selvaraj and R. Narasimhan, "Detection of sleep apnea on a per-second basis using respiratory signals," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2013, pp. 2124–2127.

[53] R. Haidar, I. Koprinska, and B. Jeffries, "Sleep apnea event detection from nasal airflow using convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process*. Springer, 2017, pp. 819–827.

[54] F. Mendonca, S. S. Mostafa, A. G. Ravelo-Garcia, F. Morgado-Dias, and T. Penzel, "A review of obstructive sleep apnea detection approaches," *IEEE J. Biomed. Health Inform.*, to be published.

[55] K. Narkiewicz, P. J. Van De Borne, R. L. Cooley, M. E. Dyken, and V. K. Somers, "Sympathetic activity in obese subjects with and without obstructive sleep apnea," *Circulation*, vol. 98, no. 8, pp. 772–776, 1998.

[56] K. Narkiewicz and V. Somers, "Sympathetic nerve activity in obstructive sleep apnoea," *Acta Physiologica Scandinavica*, vol. 177, no. 3, pp. 385–390, 2003.