Algorithms for Automated Scoring of Respiratory Events in Sleep

Rule-based models and deep neural networks

Graduation Chairman:
Prof. Dr. Ir. M.J.A.M. van Putten

Medical Supervisor:
Assoc. Prof. Dr. M.B. Westover, MD

Technical Supervisor:
Assst. Prof. Dr. E. Mos – Oppersma

Daily Supervisor:
Drs. W. Ganglberger

Process Supervisor:
Drs. B.J.C.C Hessink – Sweep

External member:
Dr. R.C.L Schuurmann

Author:
T. E. Nassi

A thesis submitted in fulfillment of the requirements of the master program:
Technical Medicine – Medical Sensing & Stimulation

May 7, 2021

Table of Contents

1 Abstract

Manual scoring of polysomnography (PSG) data, in particular respiratory event labelling, is a time-consuming task. Scoring patient recordings with highly irregular breathing and frequent apneic events is an iterative operation that may take up to multiple hours. In recent years the development and application of computer algorithms that assist manual labor has been growing immensely. Automation by such computer models has a great impact on the medical field, but interpretation of medical data is often exceptionally heterogenous. For instance, the American Academy of Sleep Medicine (AASM) provides rules for manual scoring that contain several arbitrary thresholds. This allows for dynamic interpretation of these criteria that can be manipulated between patients. In turn, this leads to increased inter-rater variability which reduces scoring consistency among annotators. Both rule-based models and machine learning algorithms offer a pallet of potentially more robust opportunities that may be applicable for automated respiratory event scoring. In this work a deep neural network and a rule-based model were designed and experimented on the worlds largest available PSG database by the Massachusetts General Hospital.

The proposed approach using a deep neural network (WaveNet) showed that a performance comparable to literature can be obtained while using a minimally invasive methodology. Differentiation between event types was possible with limited accuracy and may reflect in part the complexity of human respiratory output and some degree of arbitrariness in the clinical thresholds and criteria used during manual annotation.

Next, a completely original rule-based modelling approach to automatically score respiratory events during sleep is introduced. The AASM criteria were used as a blueprint to design a compartmentalized rule-based model architecture including hyperparameters that can be adjusted to mimic the ambiguity encapsulated in manual scoring. Global patient assessment by the model resulted in a strong agreement with the original single scorer labels. Per-event scoring led to comparable performance with current state-of-the-art models and clinical implementation opportunities seem feasible.

Preliminary results from an experiment studying the inter-rater agreement among human scorers indicates significant misclassification on event-level granularity. These findings demonstrate that new approaches should be put in relative perspective to human-to-human agreement, and not in direct contrast to single scorer data. Comparison of the inter-rater agreement between human scorers and the model showed an average decrease in Cohen's kappa value from 0.43 to 0.30. The already promising results of the proposed prototype model is expected to improve up to human-level scoring performance with future development iterations.

2 Introduction

On average every person spends about a quarter to a third of their life sleeping. This is very natural behavior, yet for most it is a mysterious state of mind. You close your eyes, lose consciousness, possibly experience some vivid dreams, and with a little luck you wake up rested and ready for the upcoming day. However, sleeping properly is not obvious for everyone, or rather, a good night rest is not possible every single night. Each of us experienced nights of little sleep, which does not feel great, both physically and mentally. You might encounter difficulties getting out of bed, and your mood may disagree with you being proactive.

Questions such as, ``What physiological processes determine sleep?" and ``What happens in case of sleep deprivation?'' have gotten significant attention the past decades and much research has been dedicated to reveal sleep's complex processes. Sleep is believed to serve a pallet of functions affecting a person's health and well-being, and impairments in sleep quantity and quality may lead to detrimental health problems, such as excessive fatigue and neurologic decline [1]. Most of these implicated functions of sleep are based on observational studies and lack the ability to establish causality, however, building upon such associations, better neurophysiologic measurements of the impact of sleep and sleep deprivation is an increasing area of research [1]. This results in exciting discoveries that continue to elucidate the underlying mechanisms of sleep and wake state coordination, paving the way for new interventions that preserve and promote optimal health.

Sleep disorders affect millions of people worldwide [2], [3]. This may lead to development of persistent fatigue which has enormous impacts on the population health [4], [5]. Accurate and timely diagnosis of a patient's sleep disorder is therefore essential. Symptoms such as sleeplessness and excessive fatigue are common, yet not specific when it comes to defining the underlying sleep disorder. Numerous disorders can lead to excessive daytime sleepiness and a thorough and detailed patient history is paramount when deciding if a patient requires laboratory evaluation [6]. Sleep disorders are divided into two classes: dyssomnias and parasomnias [2]. Both dyssomnias, abnormalities in the quantity or quality of sleep, and parasomnias, behavioral manifestations associated with the partial arousals from sleep, affect patients severely. Usually patients report having difficulties falling or remaining asleep, experiencing breathing complication during sleep or notice unusual movements during sleep [6].

Sleep apnea and related respiratory events are common types of sleep-disordered breathing that causes problems during sleep. Long periods of interrupted breathing or severe forms of obstructed respiration, often manifested as snoring, may lead to decreased oxygen intake, limiting the necessary physiological recovery during sleep. Patients experiencing such symptoms can visit a sleep laboratory where a full assessment of a person's nighttime can be performed. The gold standard to measure sleep objectively is laboratory-based polysomnography (PSG). Generally, sleep staging, respiratory events and limb movements are the three primary categories considered

while assessing patients. Experts score each of the categories to help differentiate between the various sleep disorders.

Manual scoring of PSG recordings is a time-consuming task performed by specialists in dedicated sleep centers, making this an expensive process both in time and costs. Automation of PSG analysis would decrease the required analysis time and reduce costs. Moreover, automated PSG analysis computer models could be implemented in clinical centers anywhere in the world and across a variety of data acquisition options, including home sleep testing, testing in acute care environments, specific operational conditions such as high altitude, and consumer wearable devices.

Currently at the Massachusetts General Hospital (MGH) in Boston, US, a large project aims to create algorithms that write fully automated clinical reports based on PSG data. For this report sleep staging, respiratory event scoring, and assessment of limb movement must be analyzed independently and with respect to another. Currently, automation of each individual category is studied and automated sleep staging already seems feasible [7]. Proper automated scoring of limb movement is also being researched and reasonable to good accuracy is expected, remaining respiratory event assessment as a last category that requires in-dept experimentation.

This thesis contains a description of sleep disordered breathing and the associated types of respiratory events. In addition, the current gold-standard of respiratory event scoring is evaluated and automation by algorithms is discussed. Two computer modelling methods to automatically identify apnea, hypopnea and RERA are proposed and tested on world's largest available PSG database. Finally, as an alternative to conventional single-scorer comparison, an experiment on inter-rater agreement is provided.

3 Background

3.1 Sleep disorders

Sleep-disordered breathing refers to cyclical stagnation of breathing (apneas) or reduction in inspiratory airflow amplitude (hypopneas) that leads to arterial hypoxemia or hypercapnia. These apneas and hypopneas often result in transient arousals from sleep and sleep state fragmentations throughout the night and cause overcompensatory responses of the autonomic nervous system [8]. Both apneas and hypopneas can both be classified as obstructive or central. Obstructive meaning that upper airway occlusion occurs, while central refers to the absence or reduction of neural stimulation to the upper airway muscles. In practice, the pattern of neural output and resulting upper airway muscle activation determines the phenotype of a respiratory event [9].

Patients with apnea, especially obstructive sleep apnea, are at increased risk for traffic accidents, postoperative complications and delirium [8]. Moreover, untreated apnea is associated with arrhythmias, cardiac arrest, myocardial infraction, unplanned reintubation, pulmonary embolism and pneumoniae [10], [11]. Studies that measure the apnea-hypopnea index show that an estimated 49.7% of male and 23.4% of female adults have moderate-to-severe sleep-disordered breathing, though a lower percentage are clinically symptomatic [11].

3.1.1 Obstructive sleep apnea

Sleep has pronounced effects on the respiratory system. Experimental studies have shown decreased electrical activity in medullary inspiratory neurons with efferent output to the upper and lower respiratory muscles, reflected in decreased activity of diaphragm and dilator muscles of the upper airway [8]. This can cause the tongue to fall backwards at the onset of sleep, which may in turn cause upper airway obstruction. In particular, individuals with altered mechanical properties of the upper airway are prone to local obstruction, though, anatomical processes such as alterations in craniofacial structures, enlarged tonsils, upper airway edema, decreased lung volume, and most importantly, obesity may have significant effect on upper airway airflow [9].

Obstructive apnea is strongly associated with many forms of cardiovascular morbidity and mortality [8]. And not surprisingly, weight loss and exercise are encouraged when treating obstructive apnea. Continuous positive airway pressure (CPAP) therapy is often recommended for patients with mild to severe OSA, especially in case of coexisting hypertension [9]. Other possible treatments include oral appliances, sleep position training, hypoglossal stimulation, and jaw surgery.

3.1.2 Central sleep apnea

When ventilation stops due to limited neural drive of nerves that innervate inspiratory muscles, it may result in central apneic events. During the transition from awake to sleep, cortical regulation of respiration diminishes, and metabolic regulation remains the main controlling ventilatory mechanism [12]. Arterial pressure of carbon dioxide ($PaCO_2$) plays a large role in this purely metabolic mechanism that regulates breathing, and disturbance may cause central sleep apnea. Lowering of the $PaCO_2$ in case of heart failure and other hypoxic conditions causes cessation of breathing until the $PaCO_2$ rises above the apneic threshold [13].

In contrast to obstructive sleep apnea, CPAP is only beneficial in about half of the patients suffering from central apnea [14], [15]. The remaining fraction of patients may endure unfavorable effects of CPAP when central apnea is not suppressed by medications, i.e., theophylline and acetazolamide [16]. Supplemental nasal oxygen might be helpful to stabilize $PaCO_2$ levels and therefore a patient's breathing pattern.

3.1.3 Cheyne-Stokes respirations

The occurrence and severity of central apnea highly depends on the underlying autoregulated ventilatory closed-loop regulation. Disturbances of this metabolic control may cause recurrent ventilatory overshoot and undershoot, so called "periodic breathing" [17]. Named after researchers John Cheyne and William Stokes, the Cheyne-Stokes respirations are characterized by a crescendo-decrescendo pattern of respiration between central apneic events. The magnitude of increased ventilation and decreased ventilation may amplify and continue for periods of 45 to 90 seconds [12].

Cheyne-Stokes respirations are highly correlated to central apneic events, but rarer, and the exact prevalence in the general public is unknown [12]. Since Cheyne-Stokes is believed to be a resulting pathophysiological phenomenon of heart failure and therefore associated with sudden cardiac death [17], [18].

3.1.4 Hypopnea

Apneic events vary in severity and a grey area lays between an apnea event and regular breathing. Rather than complete closure of the upper airways due to obstruction, partial narrowing and periods of hypoventilation may appear. Even so, limited neural drive to the inspiratory muscles rather than complete absence of muscular activity is possible. When an apnea-like event occurs that causes a drop in arterial oxyhemoglobin saturation and or an electroencephalographic (EEG) arousal but does not meet the criteria associated with apnea, such an event is described as a hypopnea. Hypopneas may in turn be differentiated in obstructive and central hypopneas, yet are generally grouped during scoring since differentiation is oftentimes difficult. Together with their

considered more severe apnea events, hypopneas contribute to daytime somnolence and fatigue [9].

3.1.5 Apnea hypopnea index

In clinical practice, obstructive apnea, central apnea, and hypopnea events are grouped together to compute an apnea-hypopnea index (AHI). This index describes the number of apnea and hypopnea events per hour of sleep and indicates the severity of sleep apnea in patients using the following criteria:

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

3.1.6 Respiratory effort related arousals

Even more moderate forms of sleep disordered breathing may still have negative effects on a person's sleep health. Together with a brief change in sleep state or arousal, nonhypopneic events such as respiratory effort related arousals (RERAs) are characterized by increased activation of the respiratory muscles without concomitant oxygen desaturation [19]. Such events may induce both cortical and autonomic arousals that can be illustrated by EEG measurements. Such events are believed to have milder yet sometimes similar effects on sleep fragmentation when compared to apnea and hypopnea events [20], [21]. There is no consensus on the clinical relevance of RERAs with respect to the more adverse apneas and hypopneas, but most clinics do score RERA events as it does provide additional information on a patient's sleep quality. Next to the AHI, a common index used to quantify sleep fragmentation is the respiratory disturbance index (RDI). This is defined by the number of apneas, hypopneas and RERAs divided by the hours of sleep.

3.2 Automation and algorithms

The past decades rule-based computer algorithms have assisted in iterative processes, saving great amounts of time for its users. Automation of tasks usually performed by medical staff is a development with a great impact on the medical field. Interpretation of medical data, however, can be complicated. It often requires many variables and context which is difficult to encompass by programs based on a set of rules. In recent years the development and application of machine learning has been growing immensely. This scientific study of algorithms and statistical models relying on patterns and inference rather than explicit instructions is affecting many disciplines. Especially deep neural networks, which are algorithms that can learn extremely intricate relationships between features and labels from huge amounts of data. Implementing neural networks has become very relevant in analyzing the heterogeneous kinds of data generated in modern clinical care. Many variations of neural architectures are being explored. These include convolutional neural networks (CNN), recurrent neural networks (RNN), recursive neural networks and various others, each having their own specific characteristics of processing data. For instance, a CNN can obtain great performance in the recognition of features in data such as images. Typical CNN architectures are not ideal when analyzing temporal data. This domain is better exploited by RNNs. It is very important to consider the nature of the to be analyzed data when selecting a neural network design because proper design choice in network architecture is crucial regarding its overall performance. Recently, Deepmind by Google designed a new innovative deep neural network called WaveNet [22]. Its architecture resembles a typical CNN, yet the application of dilated causal convolutions creates an increased overall receptive field. This enforces the WaveNet model to handle long-range temporal data. Though originally this model was designed to synthesize speech, its characteristics appeared applicable for the analysis of other signals. In 2018 a challenge organized by the PhysioNet Computing in Cardiology aimed to detect sleep arousals from a variety of physiological signals. Considering that the winners used a modified WaveNet model led to the conclusion that this state-of-the-art CNN architecture could indeed be applied for other purposes, such as respiratory analysis [23].

The ability of deep neural networks to learn complex patterns in large amount of data can also be disadvantageous, for instance when performing error analysis. The term ‘black box’ is regularly used in deep learning approaches as it is sometimes difficult to decipher and understand underlying architectural effectiveness of deep neural networks. Using large networks with many hidden layers, such as the WaveNet model, accentuate this problem. Underfitting, inefficient training leading to inaccurate model results, and overfitting, the inability to generalize well on new data, are common complications. Moreover, when it comes to clinical data it is important to look beyond statistical performance metrics, i.e., accuracy, sensitivity, specificity. For example, systematic bias is something of great concern. Traditional rule-based algorithms such as decision trees are easier to comprehend as its behavior is typically more predictable, and thus better understandable for its users. This is particularly true for relatively simple modelling tasks.

Performing error analysis may be easier when using rule-based algorithms. For this reason, automation of simple tasks might still favor the more traditional rule-based approaches, while deep neural networks are increasingly interesting for complex tasks, e.g., high-dimensional multi-classification tasks.

3.2.1 Automated sleep apnea detection

In the last three years a significant number of papers have been published on the detection of sleep apnea with deep neural networks, as described by recent review papers [24], [25]. Finding a patient-friendly and accurate sensor or signal, especially in combination with a suitable analysis model, is clearly an ongoing area of high relevance.

Sleep apnea detection methods typically use various breathing measurements and oximetry [25]. Alternative methods using signals derived from electrocardiography have shown some promise for predicting AHI as well, although such data has an indirect relationship to the respiratory system and therefore to sleep apnea [24], [26]. This more indirect method of analyzing respiration requires additional processing and can be affected by other illnesses including heart failure and cardiac arrhythmias, rather than sleep apnea [27]. Classification of respiratory events typically requires measuring both airflow and respiratory effort signals. Using multiple physiological signals to detect sleep apnea can provide good performance [28], [29]. However, this leads to similar problems as the current gold standard; using many different sensor signals is considered uncomfortable, expensive, and time-consuming.

The fact that automation approaches in literature use different sensors and varying model architectures makes comparison with state-of-the-art models difficult, especially since the provided performance metrics oftentimes differ as well. Moreover, most state-of-the-art models only perform global assessment and therefore lack the relevant clinical information that can be attained with evaluation on a event-level granularity [29]–[35]. Some studies do perform per-event assessment, but group all events together for binary classification, by discriminating any type of respiratory event from normal breathing [27], [28], [36]–[38]. When assessing performance, the scoring granularity (classifier type) is very relevant. A low output granularity, e.g., global AHI classification, likely leads to a better performance when compared to event classification of each second. Conversely, a higher output granularity, e.g., event level scoring, may yield more clinical information.

Below, a table with current state-of-the-art models and their performance is shown. Only studies that used a large dataset , at least 96 patients, were included.

Study	Dataset Size	Signal type	Analysis model	Classifier type	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-score (%)	AUC_{ROC} (%)	AUC_{PR} (%)
[29]	10.000	Airflow, respiration chest, abdomen, oxygen saturation	RCNN	G	88.2	-	-	-	-	-	-
[27]	2100	Abdominal effort	LSTM	A/N	77.2	62.3	80.3	39.9	-	77.5	45.3
[28]	1507	Nasal airflow, abdominal and thoracic effort	CNN1D-3ch	OA/H/N	83.5	83.4	-	83.4	83.4	-	-
[36]	1507	Nasal airflow	CNN2D	A/H/N	79.8	79.9	-	79.8	79.7	-	-
[30]	545	ECG	CNN1D-LSTM MHLNN	G	79.5	77.6	80.1	-	79.1	-	-
[31]	520	Airflow	MHLNN	G	87.2	88.3	87.8	-	-	-	-
[32]	285	Voice and facial features	GMM	G	72	73	65	-	-	-	-
[33]	188	Airflow and respiratory rate variability	LR	G	72	80	59	-	-	-	-
[39]	187	Pulse oximetry	CTM	G	87	90	83	-	-	-	-
[34]	186	Breathing sounds	Binary-RF	G	86	-	-	-	-	-	-
[37]	179	Nasal pressure	CNN1D	A/H/N	96.6	81.1	98.5	87	-	-	-
[35]	120	Breathing sounds	MHLNN	G	75	-	-	-	-	-	-
[38]	100	Nasal pressure	GNN1D	OA/N	74.7	74.7	-	74.5	-	-	-
[40]	96	Nasal pressure	GMM	OA/CA/H/N	83.4	88.5	82.5	46.6	42.7	86.7	-

Analysis models: RCNN = recurrent and convolutional neural networks, LST = long short-term memory, CNN = convolutional neural network, MHLNN = multiple hidden layers neural network, GMM = gaussian mixture model, LR = logistic regression, CTM = central tendency measure, RF = random forest. Classifier types: A = apnea, H = hypopnea, N = normal, O = Obstructive, C = Central, G = global.

3.3 Research objectives

An important question remains: “What performance obtained by a model is sufficient for clinical application?” This question is not easily answered since it is unknown what sufficient performance really means. Human scorers show significant variability when labelling respiratory events and an inter-rater agreement around 85% is considered high reliability [41], [42]. Training and evaluation of new model approaches with single scorer data, which is conventional in most literature including work that uses the MGH and Sleep Heart Health Study (SHHS) dataset, cannot exceed this 80-90% efficiency range with statistical significance. The inherent inaccuracies within single scorer data restricts any advanced training process. In the realm of data science the term label noise is often used for this problem. To identify and possibly quantify the level of label noise, it important to understand how apnea labels are currently created.

The main objective of this thesis is to design a computer model architecture that has the potential to score respiratory events with clinically acceptable performance.

To achieve this objective, first, the current gold standard of scoring respiratory events, manual labelling, is researched and its limitations are highlighted. Next, a new automated scoring method using a deep neural network is introduced. And third, an innovative approach to score respiratory events with a rule-based algorithm is explained. Both models will be assessed on individual event scoring performance and the ability to globally evaluate patients by classifying the AHI. Besides comparison with proposed approaches in literature, the scoring performance will be put in perspective with the current gold standard by studying inter-rater agreement among human-scorers. While research on the deep learning approach is currently being revised by the *IEEE Journal of Biomedical and Health Informatics*, in-depth analysis on the rule-based model is the main focus for thesis. Rules by the American Academy of Sleep Medicine (AASM) were used as a blueprint for the design of the model architecture which leads to unique clinical opportunities. Elucidation of the according advantages, and disadvantages, is included in this work.

4 Assessment of the gold standard

4.1 Scoring of respiratory events

As mentioned earlier, the MGH dataset is annotated according to a set of rules defined by the AASM. These criteria are used by technicians to systematically score apnea, hypopnea, and RERA events using a minimum of three signals including oronasal thermal airflow, oxygen saturation, and respiratory effort via respiratory inductance plethysmography (RIP), together with EEG arousals. In the upcoming sections the rules by the AASM are detailed and evaluated.

4.1.1 Apneas

- Score a respiratory event as an apnea when BOTH of the following criteria are met:
  - I. There is a drop in peak signal excursion by ≥ 90% of pre-event baseline (flow limitation) using an oronasal thermal sensor.
  - II. The duration of the ≥ 90% drop is ≥ 10 seconds.
- Score an apnea as obstructive if the flow limitation is associated with continued or increased inspiratory effort throughout the entire period of absent airflow.
- Score an apnea as central if the flow limitation is associated with absent inspiratory effort throughout the entire period of the flow limitation.
- Score an apnea as mixed if the flow limitation is associated with absent inspiratory effort in the initial portion of the event, followed by resumption of inspiratory effort in the second part of the event. *(Scoring of mixed apneas is considered optional)*

4.1.2 Hypopneas

Scoring hypopneas are defined by two rules, a recommended rule (**III. A**) and an acceptable rule (**III. B**). According to the AASM scorers can employ either one of the rule options, commonly based on varying insurance policies by different sleep centers. In practice this often leads to ambiguity when scoring hypopneas.

- Score a respiratory event as a hypopnea when ALL the following criteria are met:

  I. There is a drop in peak signal excursion by ≥ 30% of pre-event baseline (flow limitation) using an oronasal thermal sensor.

  II. The duration of the ≥ 30% drop is ≥ 10 seconds.

  III. **A)** There is a 4 ≥ oxygen desaturation from pre-event baseline.
  B) There is a 3 ≥ oxygen saturation from pre-event baseline or the event is associated with an EEG arousal

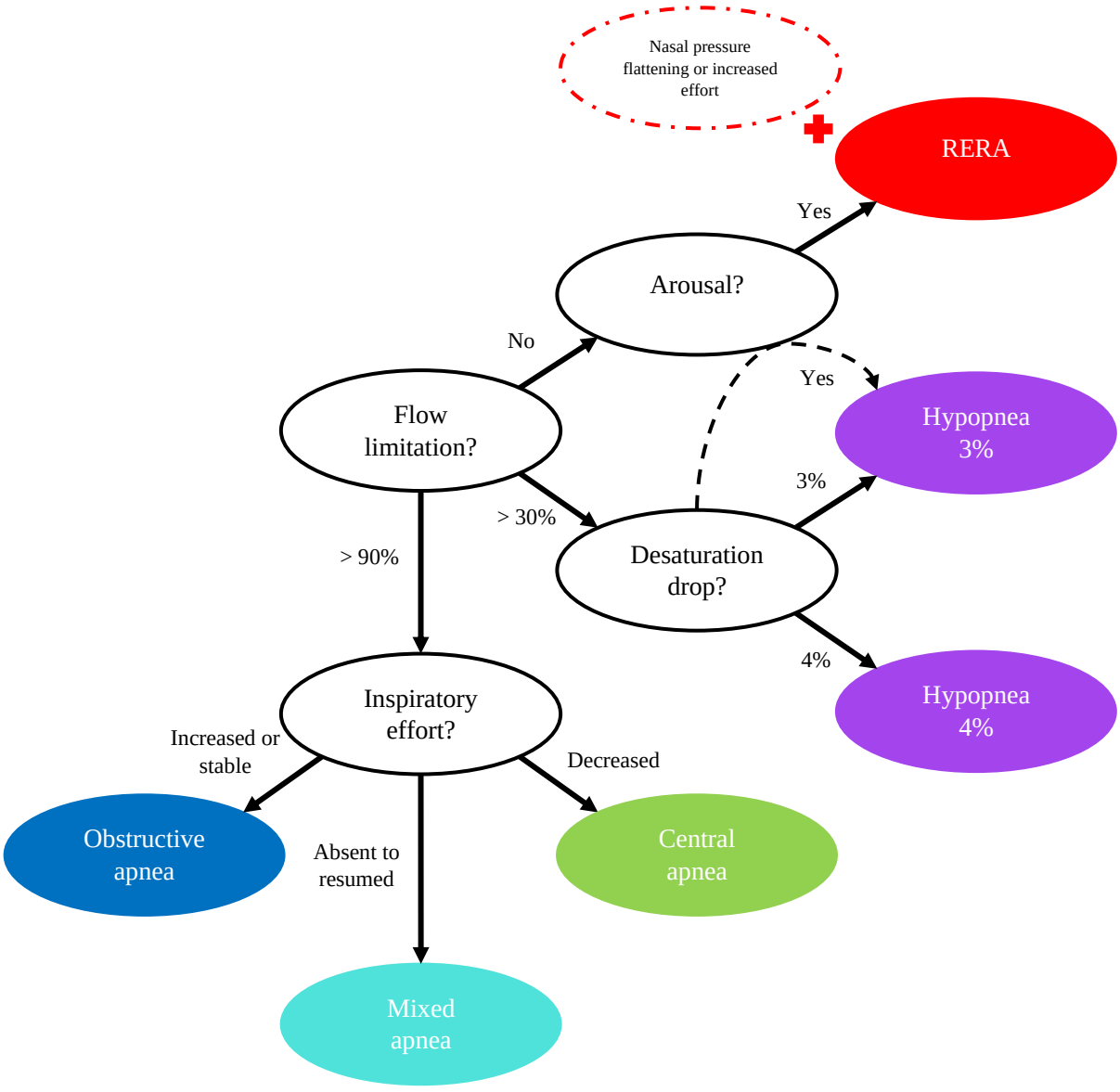
4.1.3 Respiratory effort related arousals

- Score a respiratory event as RERA if there is a sequence of breathing ≥ 10 seconds characterized by increasing effort or by flattening of the inspiratory portion of oronasal airflow in combination with an EEG arousal.

Note: When the criteria for multiple types of events are met simultaneously, the respiratory event type with the highest severity is scored. (apnea > hypopnea > RERA).

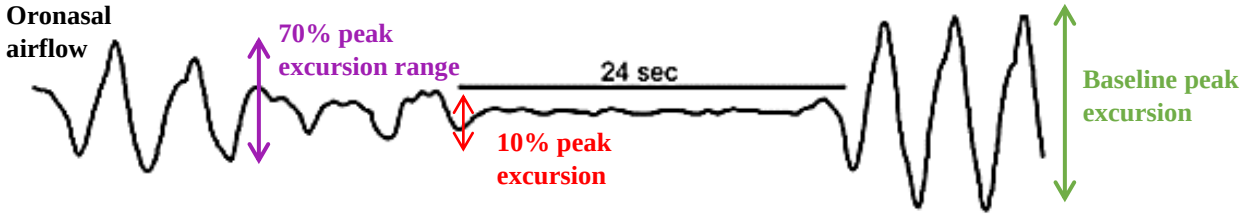
4.2 Scoring guidelines in a flowchart

The rules to score respiratory events can be interpretated as a decision tree. Below I have created a flowchart encompassing these rules.



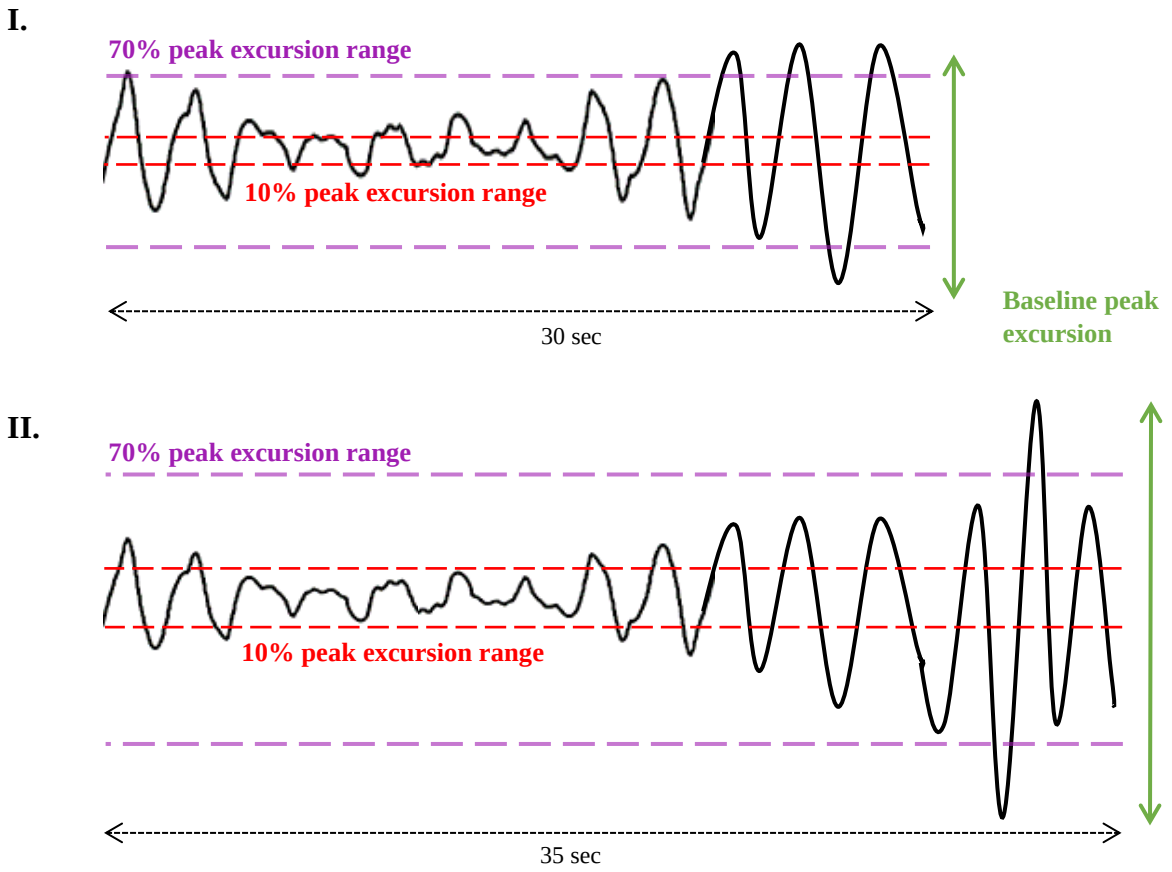
At first impression these guidelines seem straightforward. Yet, after careful assessment one might notice that there is significant room for interpretation. Each of the "white oval" questions in the flowchart embodies a disturbance in respiratory functioning. Such changes may in turn result in a measurable alteration in its represented physiological signal. However, elaborate instructions that describe what methodology must be applied to compute and quantify the level of disturbance is unaccounted for. Besides, respiratory signals measured in different patients is highly heterogeneous, meaning that one set of rules is unlikely to be suitable for everyone. In the sections below examples of potential ambiguity during scoring are illustrated.

4.3 Ambiguity when assessing flow limitation

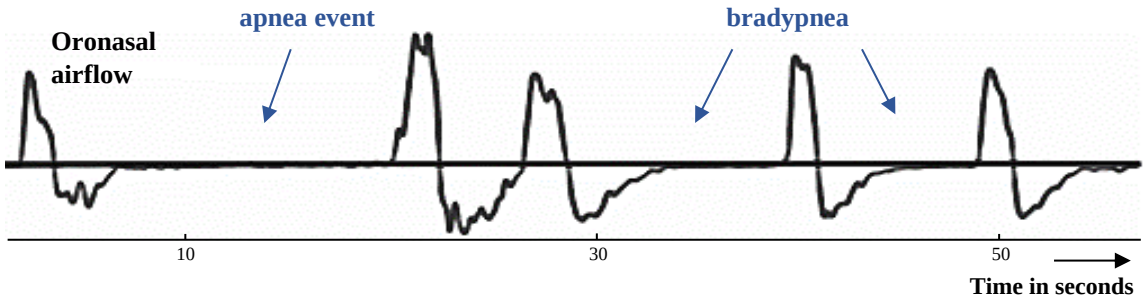


In the above figure a schematic signal segment of a patient's oronasal airflow is depicted. The sinusoidal waveform reflects measured inspiration (positive) and expiration (negative). This example depicts a typical flow limitation associated with apneic events. Note that 30% peak excursion drop is equivalent to a 70% peak excursion value with respect to its baseline. In the middle of the segment, peak excursion is vastly reduced, and nearly flat signal is observed. This exceeds a peak excursion drop of 90% with respect to its baseline excursion, for a duration greater than 10 seconds.

Now let us consider example **I.** and example **II.** in the figures on the next page. These figures represent identical events, however, for example **II.** 5 more seconds of context after the respiratory event is shown. In both example **I** and example **II** the baseline peak excursion is computed based on the largest peak excursion present in each segment. Example **II** shows more context after the respiratory event, and as a result, a large peak excursion is identified and used to compute a greater baseline value with respect to the segment shown in example **I**. This evidently leads to a larger 70% and 10% excursion range. When scoring the respiratory event in example **I**, a hypopnea would be assigned as the signal remains within the 70% range for >10 seconds, while in example **II** an apnea would be scored since the signal does not only remain within the 70% range for >10 seconds, but also within the 10% range. Comparison of both examples shows that a slight difference in computing method can lead to a different resulting annotation. This is a conceptual problem which shows that unspecific instructions may lead to variable diagnostic outcome.



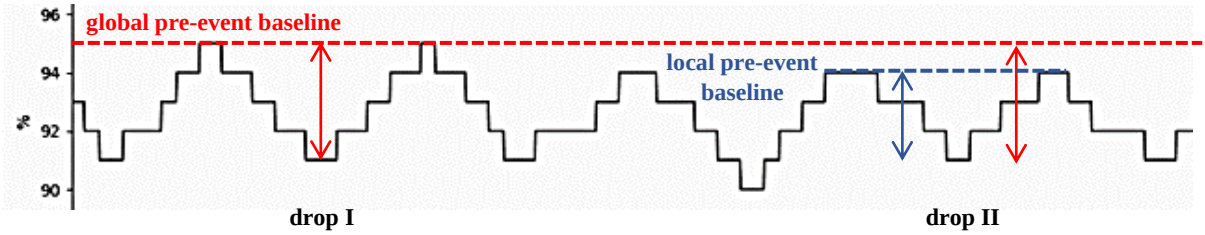
Pathophysiological ventilation by a patient may also create ambiguity during scoring. For instance, ataxic breathing or bradypnea makes for poor scoring using the rules of the AASM as the respiratory rate may be severely reduced. Increased inspiratory intervals may show very similar characteristics to short respiratory events, see figure below.



Patient recording showing oronasal airflow with an apnea event followed by bradypnea

4.4 Ambiguity when assessing desaturation drops

The next figure shows a typical oxygen saturation trace for an apneic patient. A chain of respiratory events causes recurrent arterial oxyhemoglobin desaturation followed by recovery to (near) baseline levels. All below desaturation drops are associated with an airflow limitation >30%.



When quantifying the height of each drop with respect to pre-event baseline levels, arguments can be made for multiple approaches. When looking at **drop I**, a 4% drop is observed, without much ambiguity. However, when looking at **drop II**, two options may be considered. Option A (red arrows), using a global pre-event baseline at 95%, and option B (blue arrows), using a more local pre-event baseline determined at 94%. These alternatives will lead to a 4% desaturation drop for option A, and a 3% desaturation drop for option B. Using the "acceptable" rule to score hypopneas in combination with option B will result in disregarding the hypopnea events such as associated with desaturation **drop II**.

4.5 Conclusion

The guidelines for scoring respiratory events manually have evolved over the years but remained largely driven by consensus. Thus, for example, the requirement of a 50% or 30% reduction in signal amplitude is arbitrary; there is no data to suggest that a 35% or 60% would be less ore more clinically meaningful. Moreover, visual discrimination of small percentage differences is likely poor. Besides, the differentiation of obstructive and central events is not as pathophysiologically clear as clinical scoring may suggest. Airway collapse is common during central apnea, and high-loop gain can drive obstructive events. This biological reality of blurred boundaries will be reflected in any manual or automated scoring approach.

A degree of ambiguity leaves room for interpretation while assessing a patient's respiration, as illustrated in the previous examples for the analysis of ventilatory airflow and oxygen saturation. For respiratory effort and the detection of arousals also no specifics on the extent of change in respiratory functioning are available. During scoring such subtle yet effective details matter, but are not provided by the AASM. It is, however, not self-evident that additional, more stringent, rules will improve the quality of clinical scoring, because a certain level of ambiguity does allow for dynamic interpretation of the current criteria that can be manipulated between patients.

Nonetheless, purely considering consistency during scoring a lack of details reduces robustness and is likely to increase inter-rater variability.

5 Deep neural networks for automated respiratory event scoring

This chapter details an approach in which a deep neural network can be applied to perform automated respiratory event detection. The following sections describe a shortened version of this work while an in-depth paper can be found in the Appendix.

5.1 Introduction

Recent studies show that automated apnea scoring with limited sensors (i.e., airflow or respiratory effort) can still yield acceptable performance [27], [40], [43]. Oronasal airflow measures need access to the nose/mouth, which may be difficult in specific environments. In situations where the airflow signal may not be readily acquired, an effort-belt based classification could overcome this limitation. Examples include in intensive care units, home tracking in heart failure or chronic obstructive pulmonary disease, those using nasal oxygen, and war fighter conditions. The effort belt is highly convenient, and this input signal can be acquired by a range of contact and contactless technologies in nearly every possible environment.

The ability to identify and discriminate between the specific respiratory events that are typically scored in PSG while using fewer signals is unknown to the current clinical setting. In this research we aimed to create a fully automated method that can detect respiratory events, discriminate between the different types of respiratory events, and assess the AHI with sufficient efficiency for clinical implementation using only a single respiratory effort belt.

5.2 Methods

To test whether sleep apnea prediction is feasible using minimal input information, we trained a neural network (WaveNet) to predict apneas, hypopneas, and RERAs based on a single respiratory effort signal, without the use of additional sensors that are conventional in PSG measurements. Using ~10k recordings for the MGH and ~8k recordings from the SHHS we permed respiratory event detection on global- and event-level granularity with minimal preprocessing steps. The availability of such large datasets is very valuable when training deep neural networks to prevent underfitting and restrict overfitting. 5-fold cross-validation was applied to increase the test-set size and further examine the model's ability to generalize.

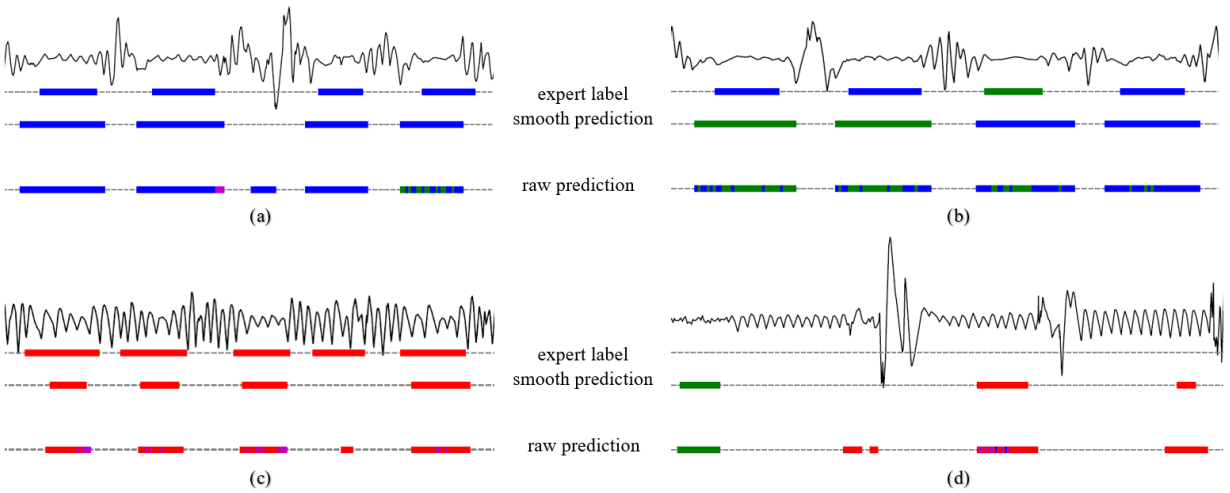
The original WaveNet architecture as defined by van Oord et al was modified by transforming convolutions from causal to non-causal and adjusting the receptive field to match our input segment size of ~7 minutes. The length of the input segments was selected to ensure sufficient context for the model to learn both spatial and temporal characteristics in the provided data. A stride of 1 seconds was used to score the complete recording with an output resolution of 1Hz. The output was later smoothed to obtain and compute a performance with event-level granularity (~18sec).

To address the large class imbalance in our data we introduced a boosted model approach by applying a binary WaveNet classifier over multiple iterations. This resulted in the removal of a large proportion of segments with regular breathing without removing many segments including respiratory events.

Next to event-level performance assessment we computed the AHI and RDI for each patient using the predicted events and the originally defined sleep duration. Next to confusion matrices, we computed the accuracy, sensitivity, specificity, precision, f1-score, and Cohen's kappa for both the event-level assessment and the global patient evaluation. Furthermore, the receiver operating characteristic (ROC) curve, the precision-recall (PR) curve and AHI correlation metrics were computed.

5.3 Results and Discussion

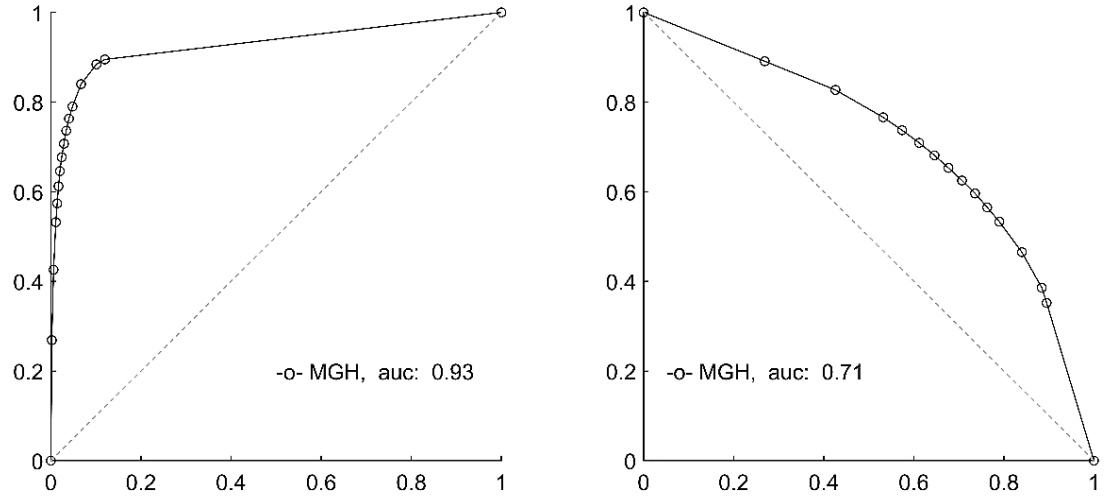
Using the methods detailed in the previous section a deep neural network method was developed to classify typical breathing disorders during sleep based on a single respiratory effort belt used in PSG. Below, some segments and associated labels are shown.



Example signal segments and the according labels and model predictions with in blue obstructive apneas, green central apneas, red RERAs and in pink hypopneas. (a), accurate predictions. (b), miss-classifications between obstructive and central apneas. (c), true positive and false negative RERA detections. (d), false positive event detections.

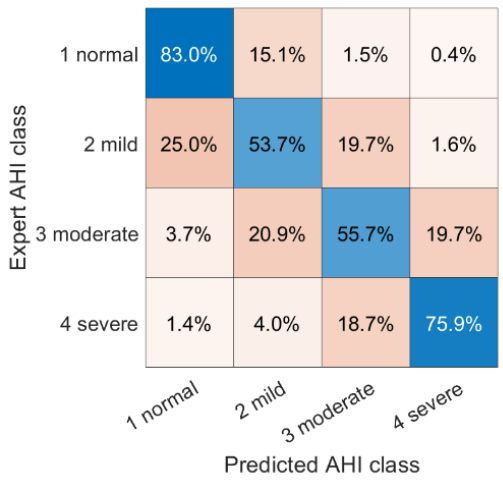
After smoothing our WaveNet model successfully discriminated respiratory events from regular breathing on the MGH dataset with an accuracy of 96%, and sensitivity, specificity, precision, F1-score, and Cohen's kappa of 68%, 98%, 65%, 67%, and 64%, respectively. Besides a high accuracy, a metric that is affected by class imbalance, the model also showed high AUC values for ROC (0.93) and PR (0.71). This means the model not only has an excellent agreement in sensitivity and specificity but also has a clinically acceptable precision in specific situations, similar to the use of home sleep apnea testing, where tolerance to especially false negatives is required [44]–

[46].We have included the F1 score and the AUCPR, as such performance metrics are not influenced by the imbalance of negative-positive classes but rather by sensitivity and precision of the positive class. The low standard deviation between the 5 folds of cross-validation (AUCROC and AUCPR mean and std of 92 ±0.5 and 71 ±1.2) emphasizes the robustness of our model on a large dataset.

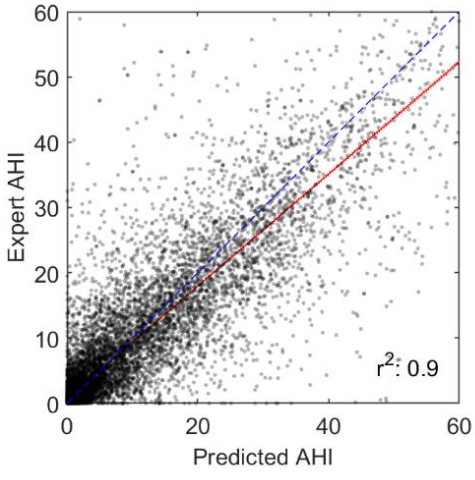


ROC and PR curves for binary classification

AHI was predicted for each patient from the number of respiratory events with an accuracy of 69%. The correlation between expert-scored AHI and algorithm-predicted AHI showed a correlation value (r^2) of 0.90. It is notable that most misclassifications of the model resulted in false positives into the neighboring AHI categories.



AHI classification confusion matrix with Cohen's kappa values of 55%



Scatter plots showing the correlation between the expert-scored AHI and the model predicted AHI

Despite decent overall accuracy, discrimination of the specific respiratory events resulted in a decreased per-event performance with respect to the first experiment. Central apneas were detected with high sensitivity of 81%, expectedly due to the apparent effect of the disorder on respiratory effort. Often markedly reduced respiratory effort is observed during central apnea events, resulting in clear features for algorithms to recognize. This is true to a lesser extent for obstructive apnea events, hence the slightly lower performance when compared to the central apneas. The recognition of hypopneas and RERAs was considered poor, with an F1-score of 29% and 31% respectively. Without additional information derived from other physiological signals the identification of hypopneas and RERAs appears difficult. It should be noted that scoring RERAs and central hypopneas are considered so difficult that the AASM scoring guidelines leaves these as “optional”, and most clinical services do not score such events. There are also several biological inconsistencies with the conventional rules for scoring central hypopneas, adding to the probability of misclassification during “gold standard” scoring.

Comparing the above performance metrics with literature (Section 3.2.1) shows that our proposed method can compete and often outperform state-of-the-art methods. Besides, discrimination between individual respiratory events while using a single respiratory effort signal is unknown in current literature.

5.4 Conclusion

The proposed method shows that a performance comparable to literature can be obtained while using a minimally invasive methodology. Differentiation between event types is possible with limited accuracy and may reflect in part the complexity of human respiratory output and some degree of arbitrariness in the clinical thresholds and criteria used during manual annotation. The use of a respiratory effort belt at the abdomen for sleep apnea analysis bears the advantage of wide implementation options ranging from acute care settings to wearable devices for home usage. Promising results were obtained in automated apnea detection with limited resources, creating new sleep assessment opportunities applicable to the clinical setting.

6 Rule-based algorithms for automated respiratory event scoring

Rule-based algorithms for respiratory event scoring has advantages over deep learning approaches, especially when it comes to performance analysis. Rule-based models often use stepwise conditional decisions to automate iterative processes. Systematic hyperparameter tuning of the underling architecture is ideal to help precisely identify the effectiveness of a given rule-based model and its encompassing decisions. Thus, evaluation of a modular algorithm may provide more insight in the underlying decisions and its effectiveness instead of considering predictive performance only. Besides, in contrast to a deep learning approach, a model based on the rules defined by the AASM will be forced to use a similar methodology as human scorers. By modelling the interpretation ambiguity present in the rules, limitations of the gold standard can be studied since manipulation of the associated hyperparameters may result in accuracy fluctuations, demonstrating limited robustness.

In the following paragraphs I explain a modelling approach to automatically score respiratory events based on the AASM scoring criteria using world's largest available PSG database. The architecture of this algorithm is modular and uses hyperparameters that can be modified to adjust for the scoring ambiguity when interpretating the required physiological signals. Modularity refers to the compartmentalized assessment of the respiratory signals, i.e., ventilatory airflow, oxygen saturation, RIP, and arousals from EEG. For this algorithm to be truly modular and allow full control for its user, all code including its design choices have been created from scratch and are completely original. Each of the three signals, i.e., airflow, oxygen saturation, and respiratory effort, are individually preprocessed, analyzed, and ultimately combined for respiratory event detection. Arousals were computed using an existing model found in literature.

6.1 Dataset description and formatting

The database used in this work was from the MGH sleep laboratory, summarized in the Table on the right. The MGH Institutional Review Board approved the retrospective analysis of the clinically acquired PSG data. Patients with and without breathing assistance by CPAP were included. In total ~13k PSG recordings were used containing diagnostic, split night and all-night CPAP PSGs.

Category		Percentage
Sex	Male: Female:	59.0 40.5
Race	American-Native: Asian: Black: Hispanic: White: Other: Unknown:	0.5 4.7 5.7 3.2 72.5 6.9 6.5
Age	< 40: 40-60: 60-80: >80: Unknown:	12.06 29.8 49.31 8.82 0.01
BMI	Under weight: Normal weight: Overweight: Obese: Unknown:	0.2 15.1 28.1 56.2 0.4
AHI	Normal: Mild: Moderate: Severe:	26.7 32.8 25.2 15.3

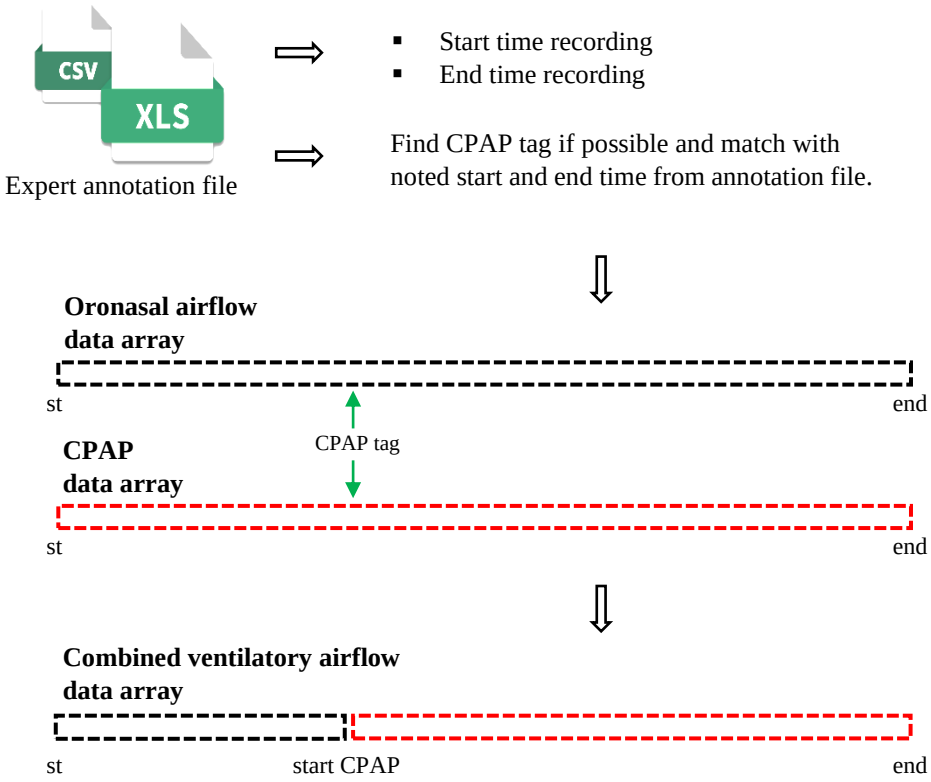
A major challenge for this work was arranging all available data into a structured and organized format. To do this automatically, customized algorithms were designed as manual formatting for a dataset of this size is not feasible. First, all relevant signals were retrieved from the online database, including:

- Ventilatory airflow, (i.e., nasal pressure or CPAP)
- Respiratory effort, (i.e., RIP at the ABD and CHEST)
- Oxygen saturation

Typically, EEG arousal computations requires an in-depth analysis of all EEG traces. For this work the EEG arousals were created using a separate model designed by the winners of the 2018 PhysioNet challenge. In this "PhysioNet model" the following 8 additional EEG signals were used to compute a continuous arousability index for the complete duration of the original PSG recording. Further details of this PhysioNet model can be found in the original study [47].

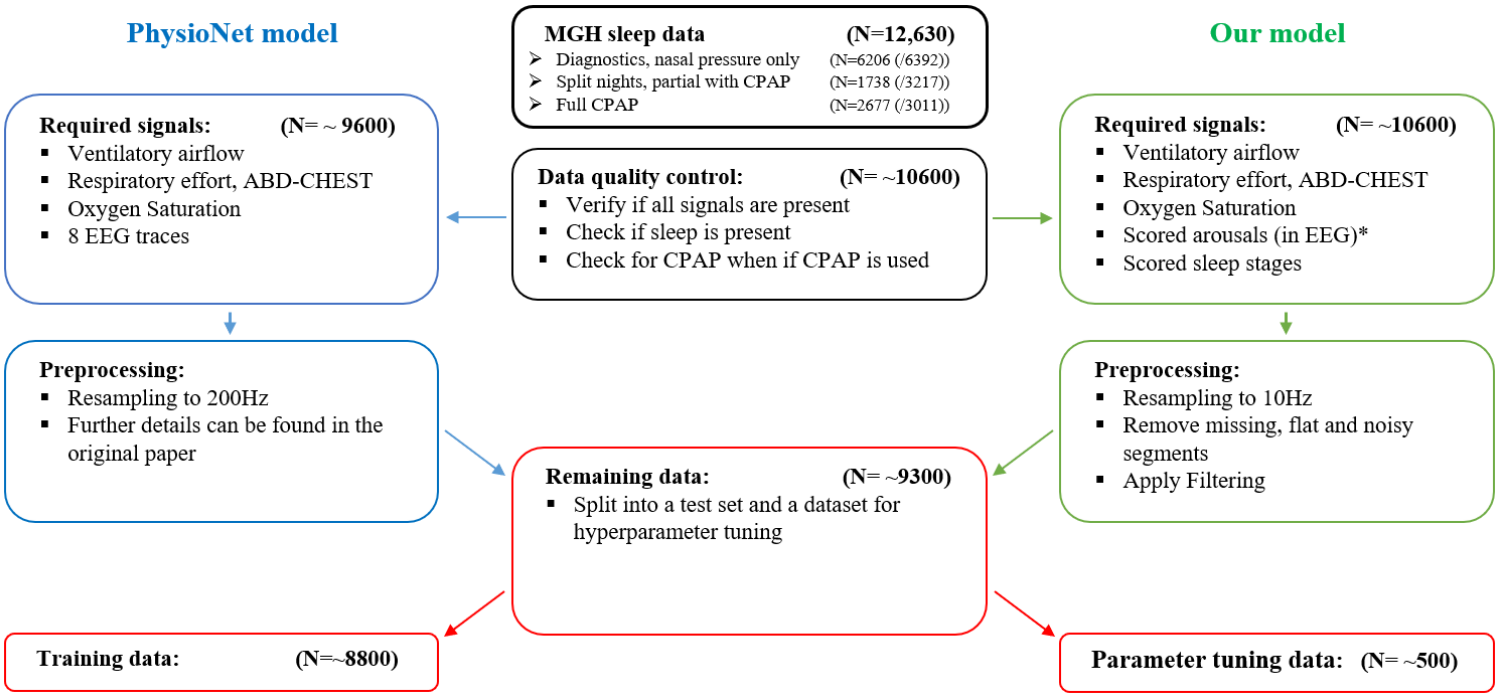
- F3_M2
- F4_M1
- C3_M2
- C4_M1
- O1_M2
- O2_M1
- E1_M2
- Chin1_Chin

Textual annotations from the experts were analyzed for tags, indicating the start and the end time of the recording, and if possible, the moment of CPAP start for recordings where CPAP was applied (all-night and split nights). When a CPAP tag was found a matching timestamp was computed relative to the start and end recording time from the expert annotation file. Next, both the nasal pressure data array and the CPAP data array were cut accordingly and combined into a single ventilatory airflow data array, see figure below. For diagnostic recordings where no CPAP is used, the full oronasal airflow data array was used.



6.1.1 Preprocessing

All signals from the MGH data consisted of a single channel with a sampling frequency of 125 Hz, 200 Hz or 250 Hz. To extract the relevant respiratory information and remove present noise for our model, minimal preprocessing techniques were applied to each signal, but oxygen saturation. A notch filter of 60 Hz was applied to reduce line noise. Consequently, a low-pass filter of 10 Hz was applied to the airflow and respiratory effort data to remove higher frequencies not of interest. For the EEG traces a low-pass filter of 20 Hz was used. Down sampling data always comes at a cost of losing valuable information. However, a low pass filter of 10 Hz used for down sampling our airflow and effort signals. This was not expected to remove significant event characteristics that limit us in identifying respiratory events, since regular breathing for adults normally ranges between approximately 0.2 - 0.3 Hz. Z-score normalization was performed using the mean and standard deviation of the 1^{st} to 99^{th} percentile clipped signal to improve data uniformity. For the PhysioNet model all signals were resampled to 200 Hz. Further preprocessing for the PhysioNet model was identical to the original work [47]. A complete workflow for the data processing is shown below.



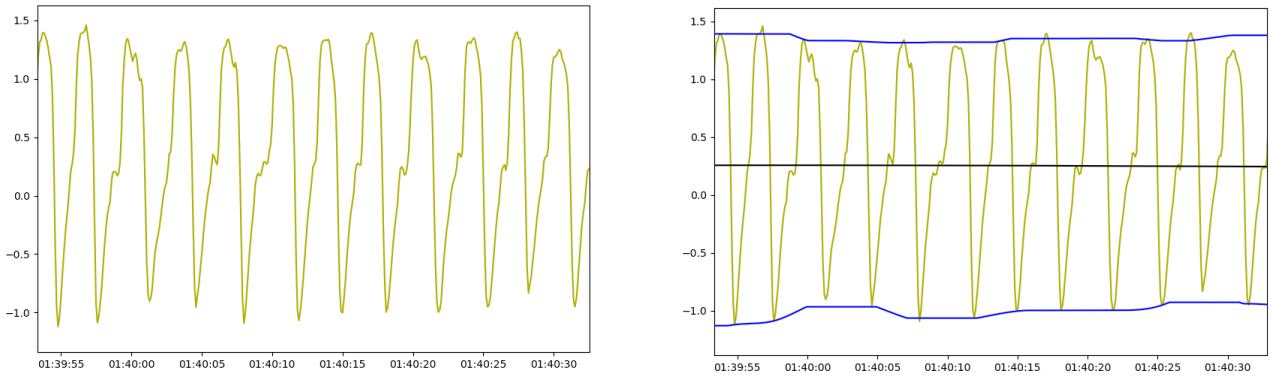
Numerous types of artifacts were identified and removed, including lost signals, missing data segments, noisy data, false annotations, movement artifacts, line noise, sampling errors, wrongly clipped data, etc. Furthermore, clinically unrealistic values were tagged and removed, e.g., oxygen saturation < 50%, Heart rates < 20 or > 250. Sleep stages were analyzed and recordings without sleep were excluded. From all successfully preprocessed data from both models (N= ~9300) a total of 500 unique patient recordings were selected for hyperparameter optimization. All other data was used for the performance analysis of our model.

6.2 Airflow analysis

Analysis of the oronasal airflow consists of the identification of unregular breathing, especially intermittent breathing. Regular breathing is characterized by a smooth sinusoidal excursion with a near stable amplitude and frequency. Among patients the amplitude and frequency may however vary. This makes any fixed characterization parameters poor in describing regular breathing for a vast and heterogeneous population. Per patients some sort of regular breathing needs to be determined to in turn identify patient specific irregularities. Such irregularities, such as apneic events, are characterized by a decrease in excursion amplitude of > 90% for a duration of > 10 sec.

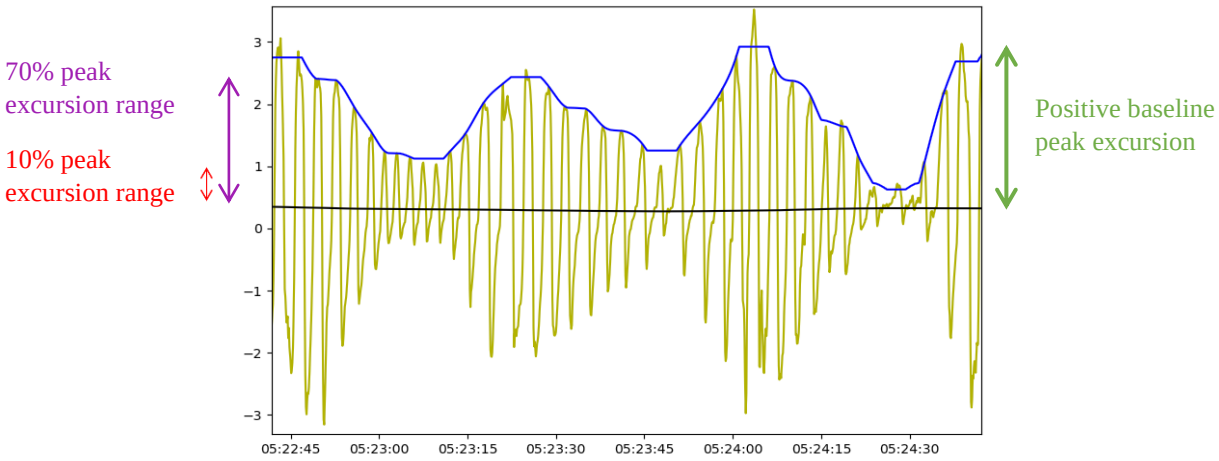
The following steps were used to assess the airflow signal:

1. To assess changes in amplitude height for the entire recording a running envelope of the positive and negative peak excursion was computed together with a ventilatory midline. An example of an airflow segment (yellow) and its determined running envelope (blue) and midline (black) can be seen below. To compute the excursion envelope first a peak detection was performed. By connecting peaks using a cubic interpolation (order=2) a continuous envelope was created. The ventilatory midline was computed by taking the mean of the positive and negative envelope. To reduce local fluctuations a moving mean with a sliding window of 10 seconds for the envelope and moving mean with a sliding window of 30 seconds for the baseline was applied.



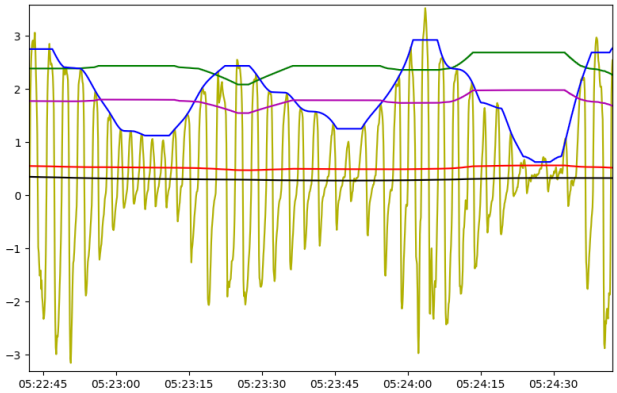
2. Using the distance between the positive excursion envelope and the midline, a positive* baseline peak excursion was computed by using a rolling quantile with a window of 30 seconds. Varying this quantile value allows for increasing and decreasing the baseline peak excursion height. This effect is most pronounced when the ventilatory envelope is unstable, e.g., very irregular breathing. By increasing the quantile values the model computes a baseline peak excursion based on the maximal amplitude of all inspirations including possible outliers (i.e., recovery breaths after a respiratory event). Reducing the quantile value will lead to a lower baseline excursion value, which results in more conservative scoring behavior. Using the baseline excursion, the 70% and 10% excursion ranges were computed to determine if

present flow limitations exceed the criteria for hypopnea and or apnea events. In the figure below the effects of reducing the quantile range from 100% to 80% when computing the baseline peak excursion can be observed. In example **I** both the 70% and the 10% excursion ranges are reduced with respect to example **II**.



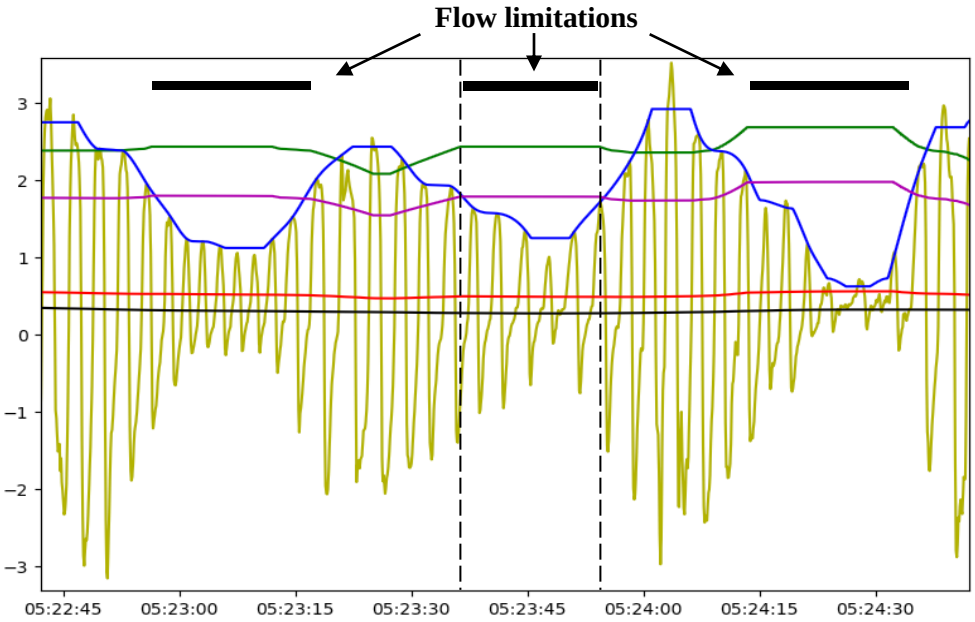
I. Peak excursion ranges computed using the 100% quantile inside a 30 second window of the airflow data.

II. Dynamic peak excursion ranges computed using the 80% quantile inside a 30 second window of the airflow data.



* Note that this slightly differs from the baseline peak excursion described in the previous chapter, in which both positive and negative excursion is considered to compute the peak excursion ranges. Future analysis will show whether using both positive and negative excursion yields better performance.

3. Apnea and hypopnea events will be identified if the airflow signal trace remains below the according peak excursion ranges for a duration > 10 seconds. In the following figure three consecutive hypopnea events can be observed, since the oronasal pressure remains below the 70% peak excursion range for approximately 25 seconds but stays above the 10% peak excursion range.



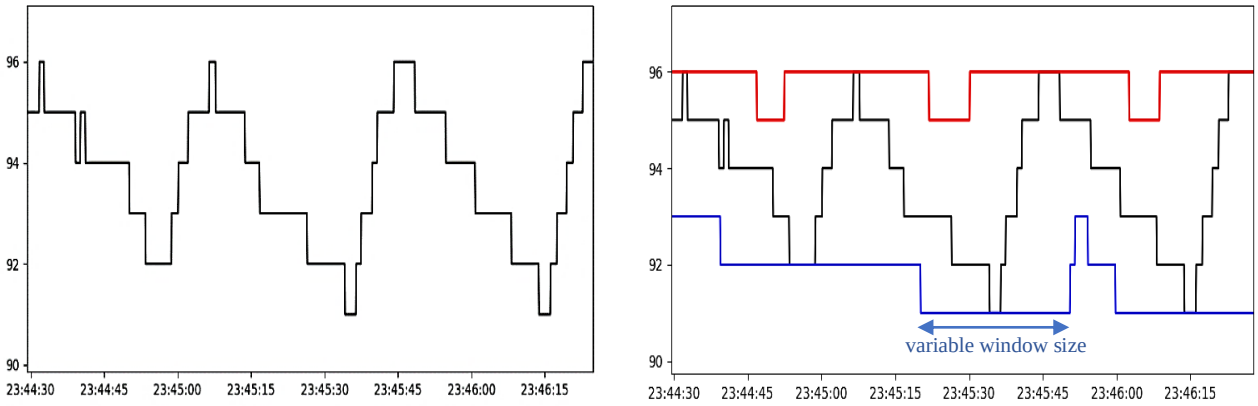
During visual assessment it may be difficult to determine the exact duration of an event. Specifically considering the ambiguity with respect to an unclear peak excursion baseline. Short lasting respiratory events with a duration around 10 seconds may therefore lead to questionable results, which is one of the uncertainties leading to increased inter-rater variability during manual scoring. In this model I therefore implemented parameters that allow varying the detection sensitivity for flow limitation duration. Increasing the minimal duration for a flow limitation will result in more conservative scoring, whereas decreasing this parameter may lead to an increment in total detected events.

6.3 Saturation analysis

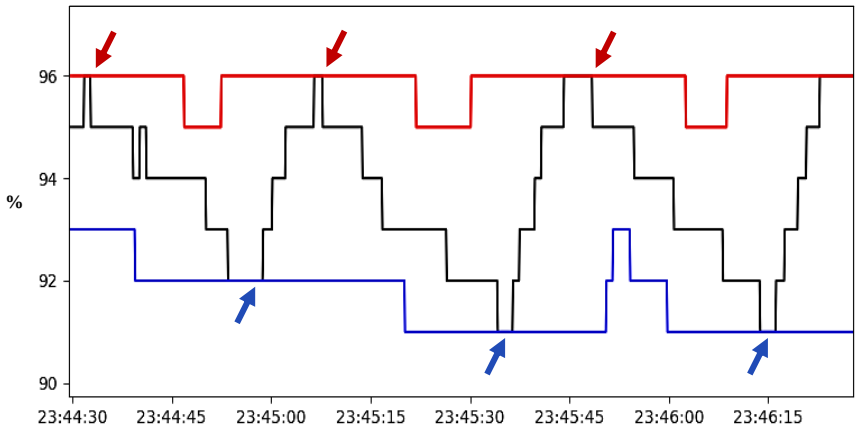
Analysis of the oxygen saturation consists of the identification of temporary saturation declines with respect to a pre-event baseline level. When a patient limits or halts ventilation a decrease in percentual oxygen-saturated hemoglobin in their blood is expected. If a patient shows a severe oronasal flow limitation, i.e., severe apneic event, with a duration of 30 seconds, a temporary decline in oxygen saturation levels is practically imminent. A patient that exhibits many severe respiratory events in sequence may even reach hypoxic levels which can be harmful for numerous reasons. However, when a flow limitation of 35% for a duration of approximately 10 seconds occurs, i.e., mild hypopnea, a drop in saturation may be limited or absent. Considering the level of saturation drop with its associated flow limitation is essential when assessing the severity of respiratory events.

The following steps are used to compute saturation drops using the 4% desaturation rule:

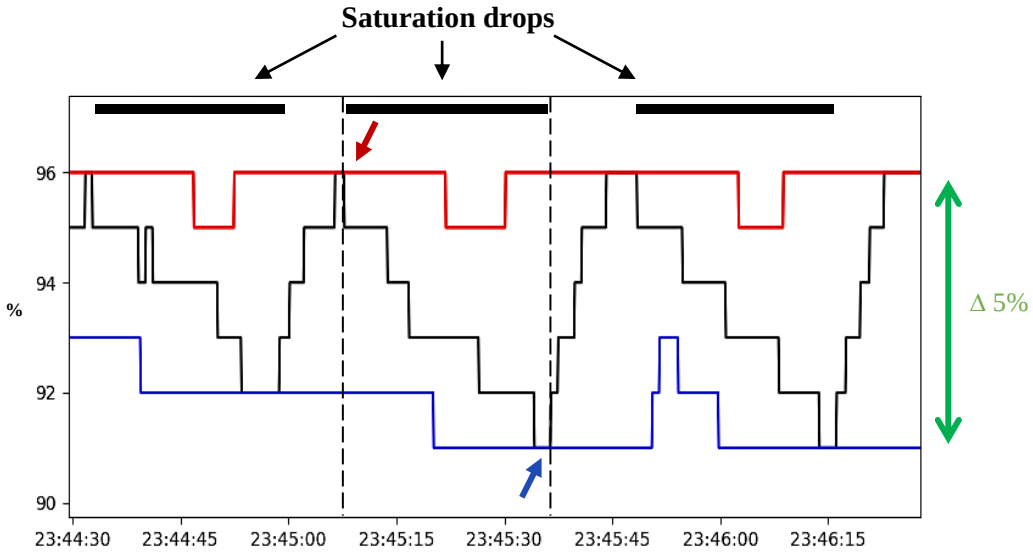
1. The rolling maximum (red) and rolling minimum (blue), with a window of variable size, of the saturation signal was used to compute pre-event baseline saturation levels and drops in saturation, respectively. Using a large sliding window (Section 6.2 step 1) caused short consecutive drops to be indistuiguisshable, whereas a small window disregarded slow decreasing drops. Analyzing the saturation trace in multiple iterations and varying the window size from 5-30 seconds will cause early iterations to detect sudden drops while later iterations will identify long slow declines in the saturation signal.



2. Whenever the saturation trace dropped below pre-event baseline represented by the rolling maximum, locations with saturation drops were tagged (red arrows). When the saturation trace started increasing from its lowest saturation level represented by the rolling minimum, possible end points for saturation drops were tagged (blue arrows).



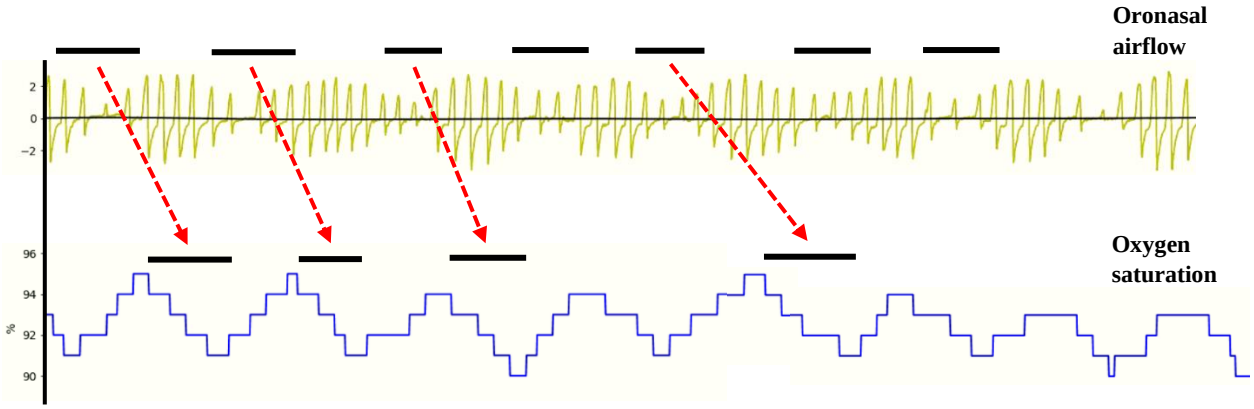
3. Matching start and end locations were grouped and the respective height of each drop was determined (green arrows). If a saturation drop was > 4%, the associated location was marked.



The described method to quantify saturation drops uses the local baseline approach as explained in Section 4.4. Multiple scorers from the MGH and the MST confirmed that this is applied most commonly.

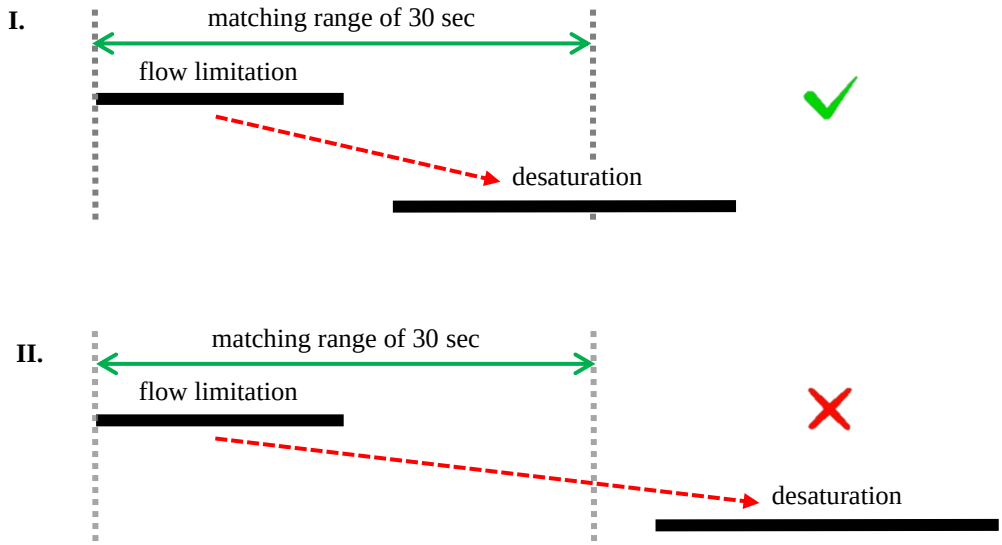
6.3.1 Matching flow limitations and desaturation drops

Halted breathing during flow limitations reduces oxygen uptake in blood, however, a drop in saturated blood percentage is not instant. Rather, delayed coexisting physiological disturbance (typically ~10 seconds) is observed when comparing saturation measurements with recorded airflow. Such shifts in time have to be accounted by the model but may become difficult for patients with events that are in rapid succession. In the example below can be observed that desaturation drops do follow flow limitations with a delay that is not consistent. Moreover, some desaturation drops show a decrease of 4% while others only a decrease of 3%, which must be handled in a different manner. The red arrows show flow limitation that are associated with a 4% desaturation drop.

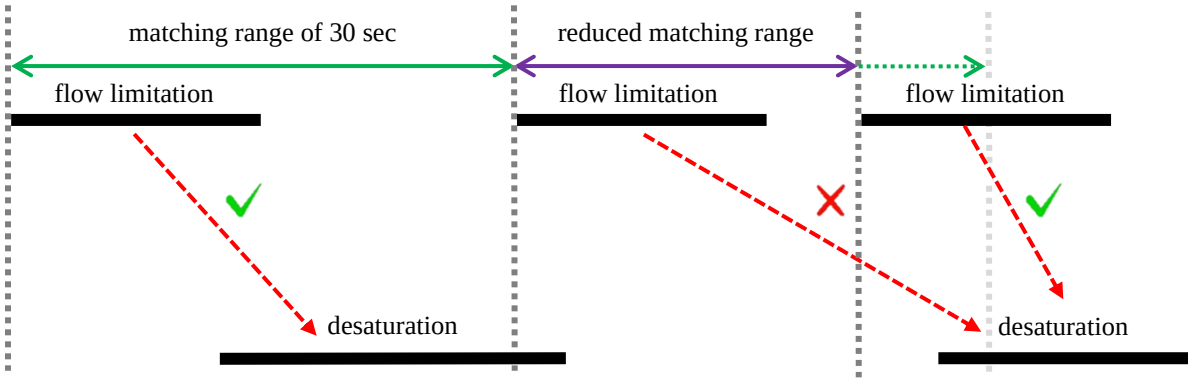


The duration of delay is difficult to quantify as the start and end location of flow limitations may be unclear during assessment. Besides, the exact start and end location are not of great importance during manual scoring, but for algorithms it may create difficulties since specific instructions are required to match events. In this work the following matching method was applied.

- Flow limitations and desaturations were matched if a detected desaturation drop commenced within 30 seconds of the start of the flow limitation, **I**. When a desaturation was found outside of the 30 second range, events were not matched **II**.

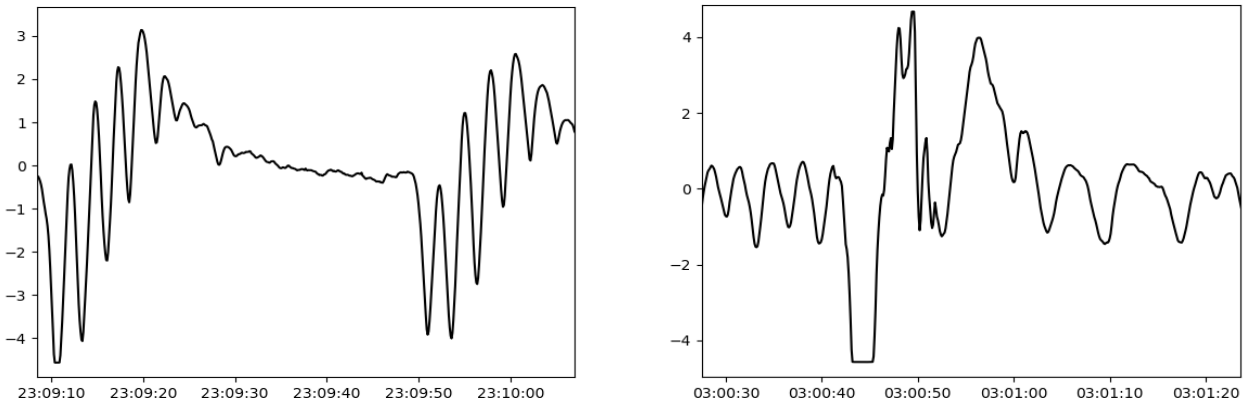


- The 30 second matching range of events was reduced to the start of a following event, if a following event commenced within the 30 range. This prevented matching of flow limitations with desaturation drops associated with neighboring events.



6.4 Respiratory effort analysis

Analysis of RIP signals for respiratory event analysis consists of the identification of increasing and decreasing signal amplitudes during flow limitations or arousals measured with EEG. When apneic events are observed in the airflow trace, characteristics in the respiratory effort signals help differentiate apnea types. Increased respiratory effort indirectly indicates increased neural drive to the respiratory muscles, often seen with obstructive events, whereas decreased respiratory effort suggests limited or absent neural innervation of the inspiratory muscles, associated with central events. Sudden magnification in amplitude of the respiratory effort signals associated with an EEG arousal is a hallmark for RERAs, if the criteria for an apnea or hypopnea event at that location are not met. Below, examples are shown of typical increased and decreased respiratory effort during apnea events.

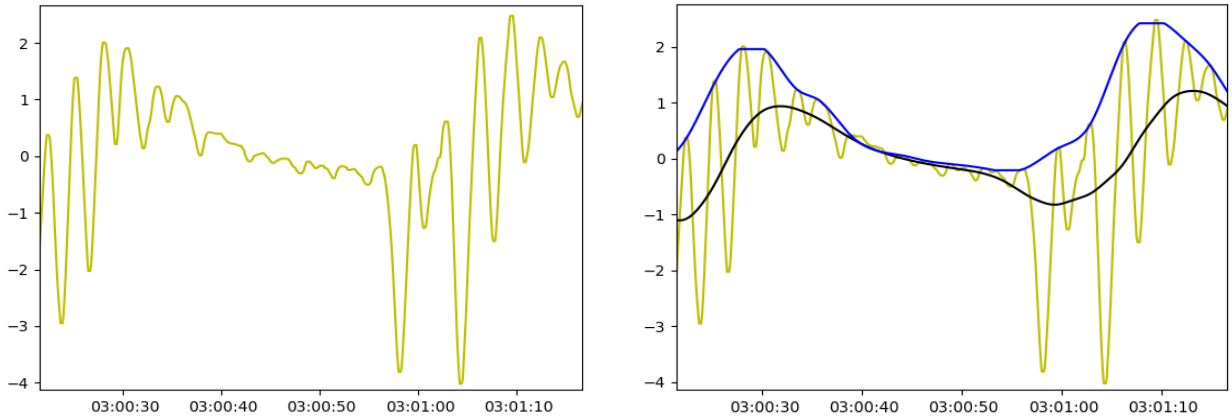


Abdominal RIP trace showing decreased respiratory effort (left) and increased respiratory effort (right) for an apneic patient.

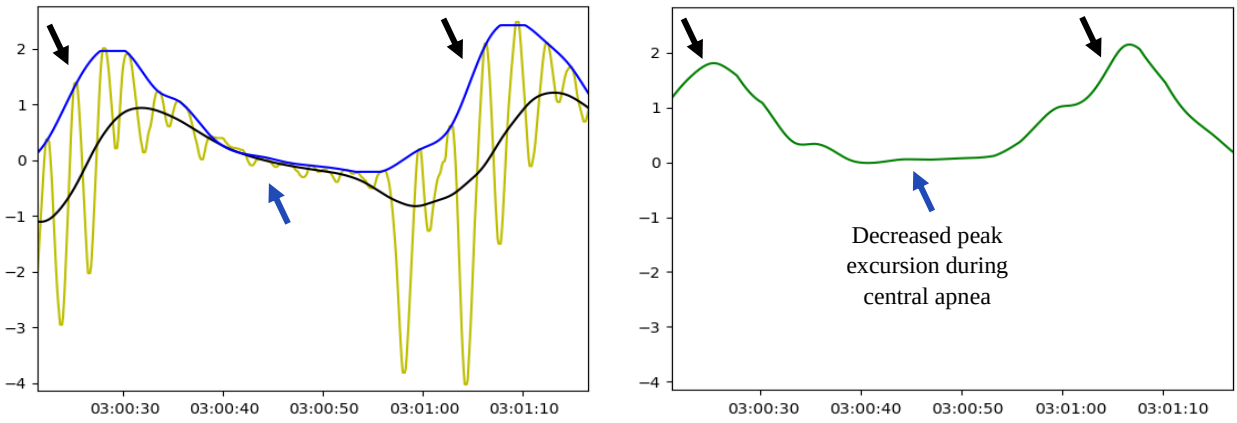
In this work the following steps were used to assess respiratory effort signals measured at the abdomen and chest area. Both signals were individually analyzed and averaged to ultimately evaluate the combined change in characteristics. In this way, less false results were expected due to poor signal quality in one of the channels. Respiratory effort traces regularly show significant baseline shifts, which may make visual analysis of peak excursion more difficult. Using a static threshold to determine changes in amplitude height will often be inaccurate when large baseline shifts are present. Therefore, a dynamic approach was used to determine local alterations in amplitude height.

1. Effort signals were smoothed using a moving median with a sliding window of 0.5 seconds to remove unrealistic high frequency fluctuations. Using this smoothed signal, a positive envelope and according midline were computed by applying the same approach as used when assessing flow limitations. In the figure below on the right can be observed how a baseline shift present in the sinusoidal effort trace is captured by the model. The midline (black) follows

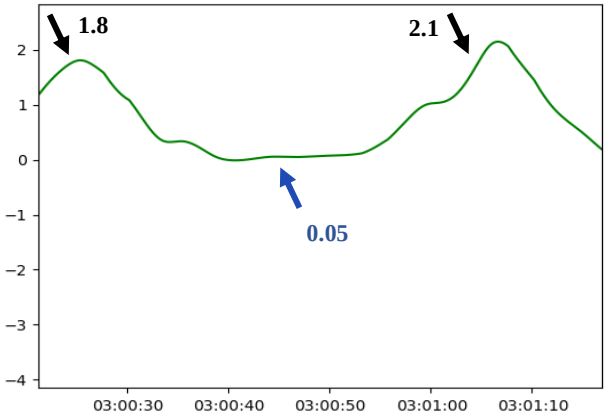
the trend of the effort trace (yellow) while the positive envelope (blue) accurately connects all peak amplitudes.



2. By computing the difference between the midline and the positive envelope, increasing, and decreasing peak excursion height can be captured dynamically. In the figures below the peak excursion height over time is visualized (green). A clear decrease in peak excursion is observed in the middle of the segment (blue arrow), indicating decreased respiratory effort for a duration of approximately 20 seconds.



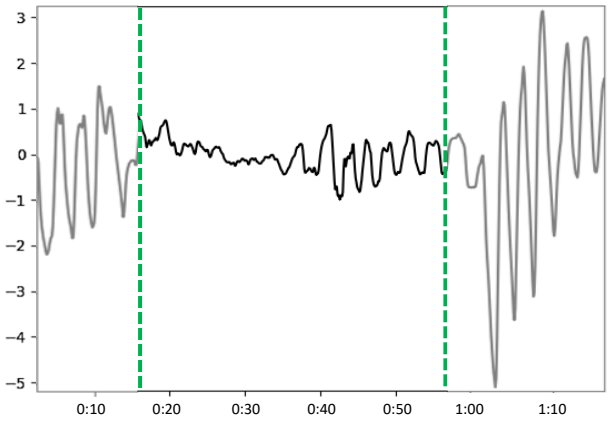
3. Using the peak height excursion measured before, after, and during an event, an inspiratory peak excursion ratio was computed. The baseline peak excursion before and after the event was averaged and divided by the peak excursion during the event, see example below. A ratio threshold to discriminate between central and obstructive apneas was 20. Flow limitations that meet the criteria associated with an apnea event that also show an inspiratory peak excursion ratio ≥ 20 were classified as central apnea. Events with an inspiratory peak excursion ratio < 20 were classified as obstructive apnea.

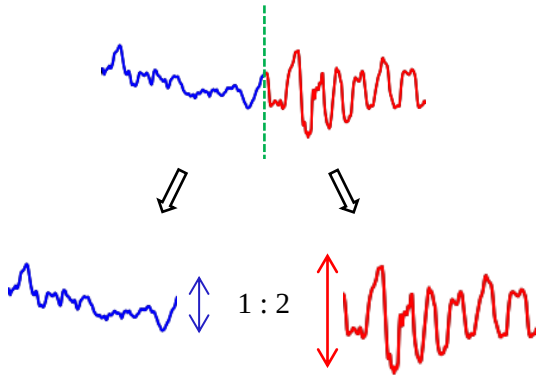


$\Rightarrow$ *peak excursion ratio*: $\frac{(1.8 + 2.1) / 2}{0.05} = 39$

- ratio ≥ 20: event is central
- ratio < 20: event is obstructive

4. Apnea events that show an inspiratory ratio of ≥ 20 are believed to encompass a central aspect, meaning that during the event limited neural drive of the respiratory muscles is present. This central aspect does not necessarily demonstrate itself for the entire duration of the apnea. Mixed apneas are characterized by an initial central component, that later develops into resumes respiratory effort. Apnea events with a central component were therefore cut in half, and consecutively compared in peak height excursion. If peak excursion of the latter part is ≥ 2 times greater than the respective leading part of the event, the model classified this event as mixed apnea.

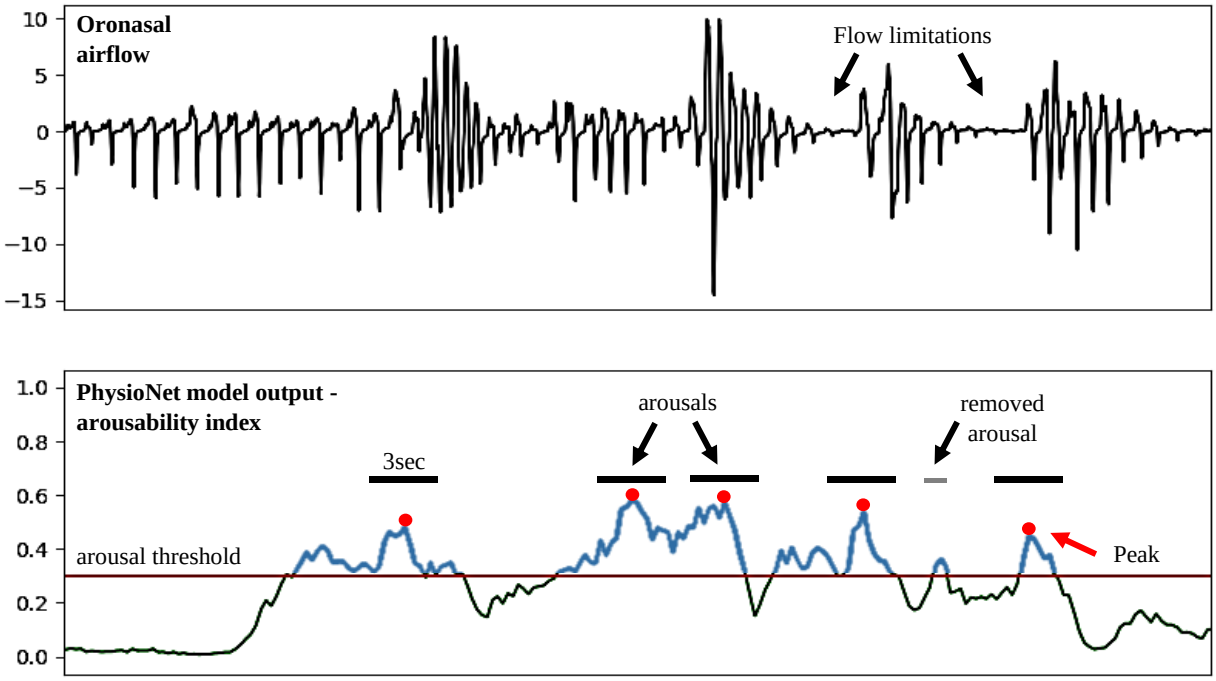




The inspiratory ratio threshold was based on ~500 scored apnea events. Ranging the ratio threshold between 0-100 the optimal discriminatory ratio value was determined by selecting the value that corresponded with the highest agreement with the expert apnea events. A similar method was used to determine the optimal excursion height value to discriminate between central and mixed apnea events.

6.5 Creating EEG arousals

Typically, human technicians use 6 EEG traces to score EEG arousals. Even though arousals are not required for scoring apnea and hypopnea events, commonly arousals do present themselves when such events are detected. Arousals are required for scoring the less severe RERA events, which are characterized by an EEG arousal in combination with either flattening of the inspiratory portion of oronasal airflow or joined with increased effort. As mentioned, in this work I used an existing model in literature that creates a continuous arousability index. A fixed threshold value was used to identify EEG arousals. For all segments exceeding the arousal threshold, peak detection was performed to find locations where the PhysioNet model resulted areas with the highest arousal conviction. Detected peaks were converted into arousals with a fixed duration of 3 seconds. In the figure below can be observed how the converted EEG arousals correspond with consecutive flow limitations in the oronasal airflow trace. By finding the highest agreement with ~500 originally scored arousal events a value of 0.3 was used as a threshold to create EEG arousals. Short events were removed, meaning that the complete segment above the arousal threshold had a duration < 3 seconds.

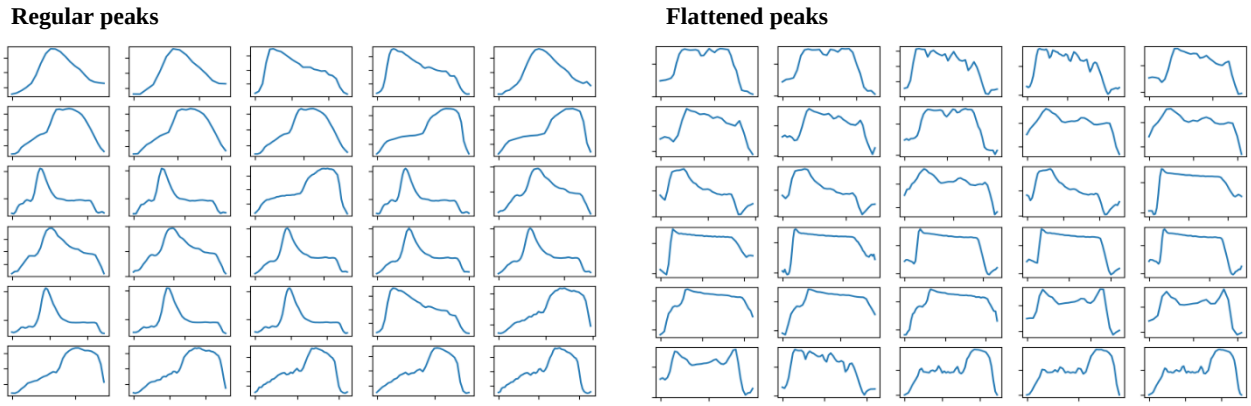


Continuous arousability index converted into separate 3 second arousals (blue) according to peak detection (black dots) in all segments exceeding the arousal threshold (blue). Comparison with the oronasal airflow, peak excursion drops seem to match the location of the EEG arousals, indicating potential areas with a respiratory event.

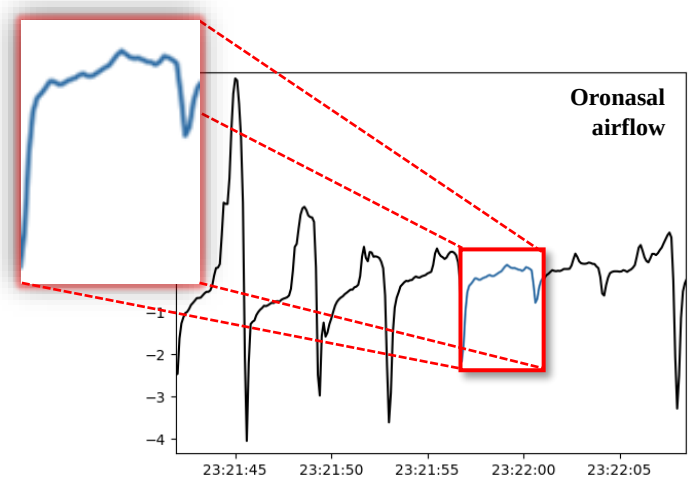
6.5.1 Flattened inspiratory peak detection

All areas with EEG arousals were assessed for possible RERA events, only if no flow limitation were found which would indicate a more severe apnea or hypopnea event. Either local increased inspiratory effort or flattening of the inspiratory portion of the airflow would be required to score RERA events. To identify flattened inspiratory peak excursion of the airflow, kernel convolution matrices were used, commonly known for image processing applications.

1. 30 real examples of regular inspiratory peaks and flattened airflow peaks and were normalized and stored. Scaled copies of the examples were added using a scaling factor of 0.75 and 1.25 to increase heterogeneity and help encompass the large variability in peaks observed in different patients. 2 separate groups were formed, 1 with regular peaks, and 1 with flattened peaks.



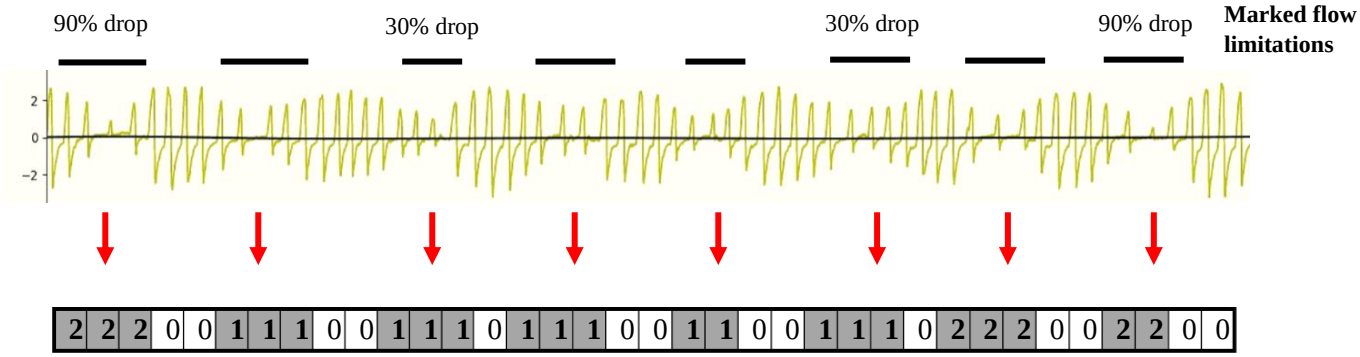
2. All example peaks were projected on the airflow trace and the group with the highest agreement determined classification of the segment as regular or flattened. To limit the effect of outliers in both peak groups, the 80% quantile among the peaks in the respective groups was used.



6.6 Event indexing and labelling

After analysis of the individual signal traces the 4 signals were evaluated together, relative to time. For computational assessment of simultaneous disturbance, areas with flow limitations, desaturation drops, changes in respiratory effort and EEG arousals were using the following steps:

1. For each signal all marked locations were converted into a 1-dimensional data array. These data arrays were constructed so that a single array corresponded with a full recording by using the same sampling frequency of 10 Hz. Thus, each index of these arrays related to one datapoint in the corresponding signal. In the figure below and schematic visualization is shown where flow limitations derived from an oronasal airflow signal were converted using the according indexing table.

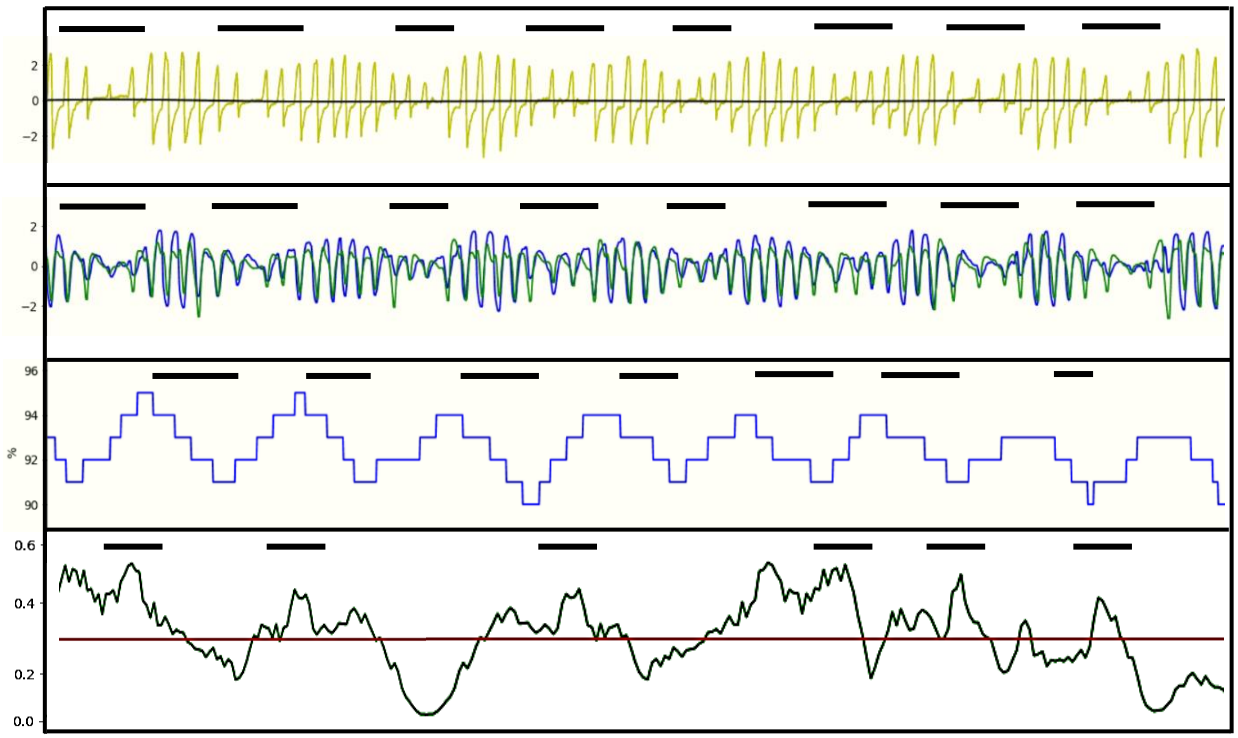


Indexing Table:

Signal trace	index == 0	index == 1	index == 2
Flow limitation	no	30% drop	90% drop

2. Stacking the indexed version of the four signals resulted in a 2-dimensional representation of a complete PSG recording. A complete schematic figure of the indexing approach is displayed on the next page.

For computational assessment of coexisting physiological disturbance, the 4 signals were stacked. All event locations and the according type of disturbance was extracted.

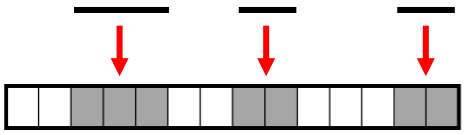


⇩

Indexing Table:

Signal trace	index == 0	index == 1	index == 2	index == 3
Flow limitation	regular breathing	30% drop	90% drop	
Effort change	stable	increased	decreased	decreased → resumed
Desaturation drops	no drop	3% drop	≥ 4% drop	
EEG arousals	no arousal	arousal		

Data arrays were constructed with all event types and locations for each full PSG recording.



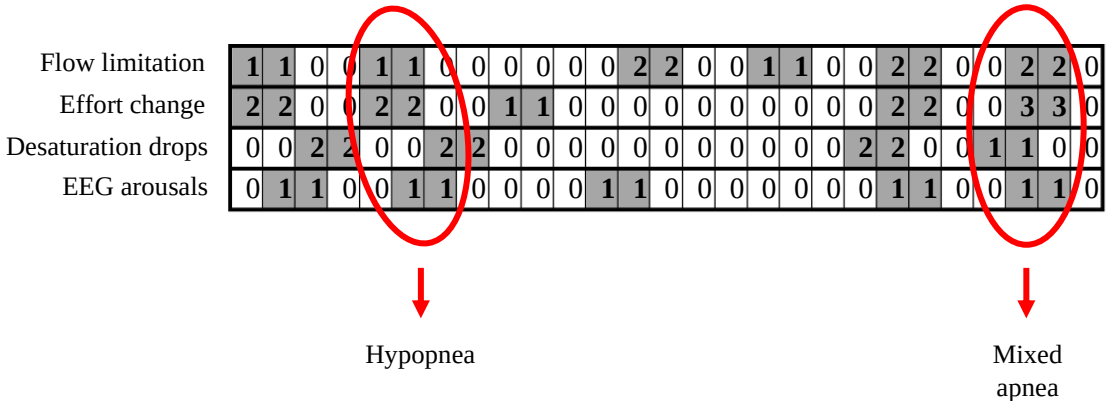
⇩

Resulting in stacked 2-dimensional structures with all events of the 4 signals

Flow limitation	**1**	**1**	0	0	**1**	**1**	0	0	0	0	0	0	**2**	**2**	0	0	**1**	**1**	0	0	**2**	**2**	0	0	**2**	**2**	0
Effort change	**2**	**2**	0	0	**2**	**2**	0	0	**1**	**1**	0	0	0	0	0	0	0	0	0	0	**2**	**2**	0	0	**3**	**3**	0
Desaturation drops	0	0	**2**	**2**	0	0	**2**	**2**	0	0	0	0	0	0	0	0	0	0	0	**2**	**2**	0	0	**1**	**1**	0	0
EEG arousals	0	**1**	**1**	0	0	**1**	**1**	0	0	0	0	**1**	**1**	0	0	0	0	0	0	0	**1**	**1**	0	0	**1**	**1**	0

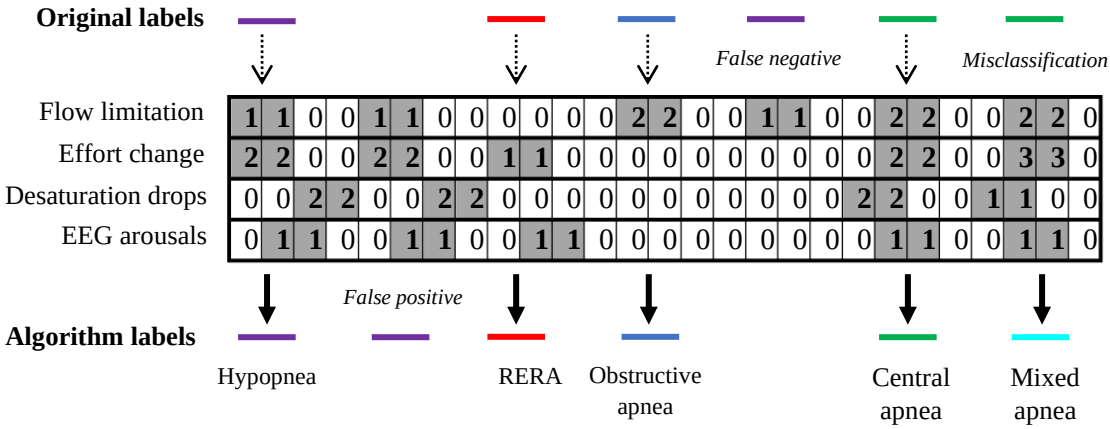
3. The organized structures allowed for systematic labelling of the respiratory events using the AASM scoring rules. Apneas and hypopneas were scored based on the locations of found flow limitations. Associated change in respiratory effort dictated the type of apnea, whereas the height of desaturation drops defined the type of hypopnea. EEG arousals matched with airflow disturbance or increased effort respiratory effort led to the detection of either hypopneas or RERAs.

    - All flow limitations with a ≥ 90% excursion drop were scored as apnea. To differentiate between the types of apnea, i.e., obstructive, central, or mixed, change in effort was evaluated. If simultaneous stable or decreased effort was found, an obstructive apnea or central apnea, respectively, was scored. When decreased effort was found in the initial part of the flow limitation, that later developed into resumed respiratory effort, a mixed apnea was scored. (see Section 4.1.1)
    - Flow limitations that did not meet apneic thresholds were matched with desaturation drops. Events matching with a drop ≥ 4% resulted in the identification of a hypopnea using the acceptable rule (**III.A**). Also including events associated with 3% desaturation drops or matching arousals lead to the detection of hypopneas using the recommended rule (**III.B**). (see Section 4.1.2)
    - All areas with EEG arousals were evaluated if the criteria for apnea and hypopnea at that are were not met. When matching increased inspiratory effort or flattening of the inspiratory portion of the airflow was found, the associated area was labelled as RERA. (see Section 4.1.3)



6.7 Model evaluation

After analysis of the signals and the labelling of their indexed representation, the rule-based algorithm was evaluated by comparison with the original labels. The original labels were projected on the signal traces, also in 10 Hz granularity. Sleep and wake time for the patients was computed using the already scored sleep stages from the original annotations. All events detected during wake time for patients were disregarded.

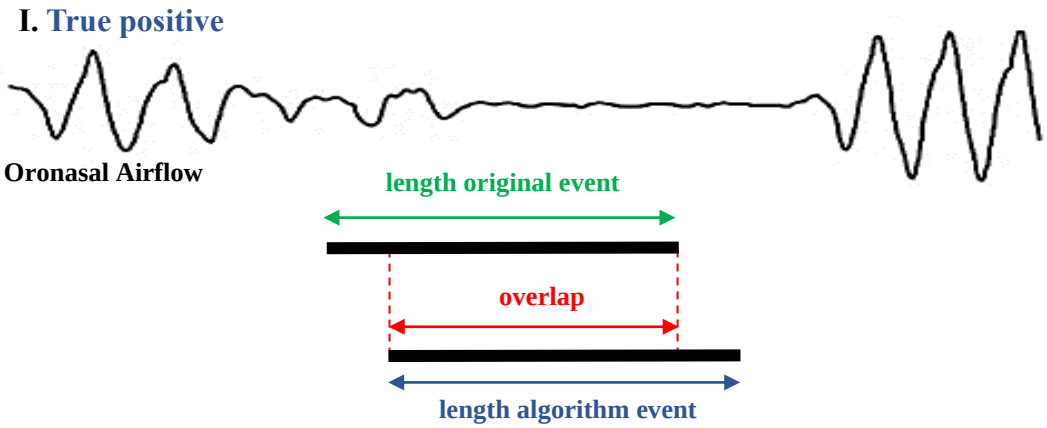


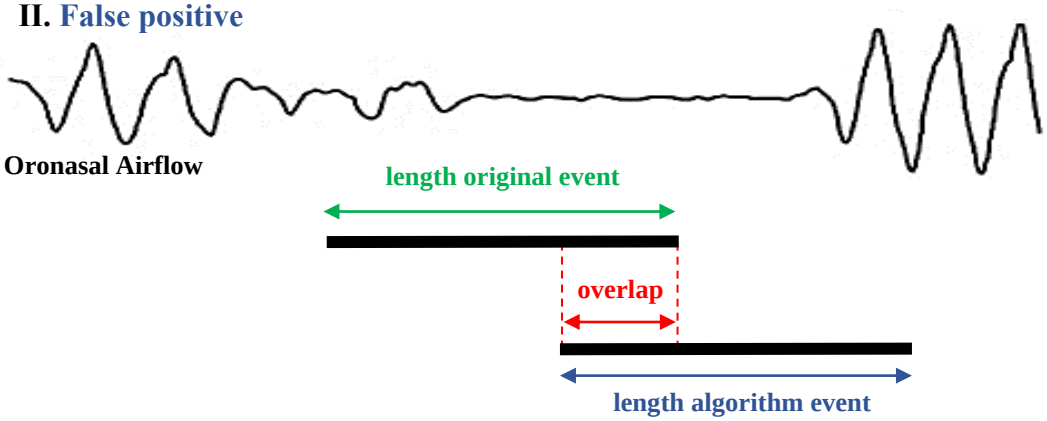
6.7.1 Per-event evaluation

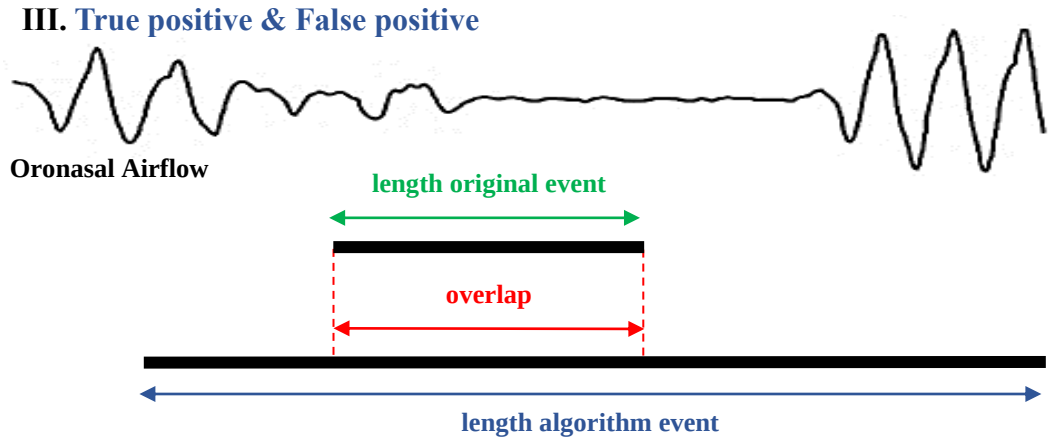
Since the exact start and end location of respiratory events is not all too meaningful, an algorithm event was considered correct when more than 50% of its duration overlaps with an expert label. Confusion matrices were computed to assess per-event performance of our algorithm. The true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) were computed for each of the 6 classes (5 respiratory event types and a 'no-event' class).

In the following figure three evaluation examples of algorithm events are shown. In situation **I.** the length of the overlap is greater than the 50% of the original event, therefore regarded as a TP. In situation **II.** the length of the overlap is smaller than the 50% of the original event, therefore regarded as a FP. In situation **III.** is a more complex. Similar to the first example a TP is found. Additionally, a FP is found because the length of the algorithm event is >2 times the length of the original event.

Situation **III.** would indicate that the algorithm was extremely sensitive to an extend that that is multiple events could have been found in this location. By using the 50% rule we therefore both complimented and penalized the performance of the algorithm.


I. True positive
Oronasal Airflow
length original event
overlap
length algorithm event


II. False positive
Oronasal Airflow
length original event
overlap
length algorithm event


III. True positive & False positive
Oronasal Airflow
length original event
overlap
length algorithm event

The TP, TN, FP, and FN values were used to determine the following event-per-event performance metrics of the algorithm.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad Sensitivity = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP} \qquad Specificity = \frac{TN}{TN + FP}$$

$$F1\ score = \frac{2x\ Sensitivity * Precision}{Sensitivity + Precision}$$

Additionally, Cohen's kappa values were determined using the formula in Cohen's original work.

6.7.2 Global evaluation

In addition to event-per-event evaluation the global scoring performance was assessed. Global assessment of sleep apnea severity is typically used for clinical diagnosis [24]. We determined the AHI and RDI value per patient using the following equations.

$$AHI = \frac{Obstrive\ apneas + Central\ apneas + Hypopneas}{hours\ of\ sleep}$$

$$RDI = \frac{Obstrive\ apneas + Central\ apneas + Hypopneas + RERAs}{hours\ of\ sleep}$$

With the AHI score all patients were categorized as normal, mild, moderate, or severe sleep apnea. Categorization was according to the conventional criteria as defined by the AASM (see 3.1.4). The classification accuracy was obtained by creating a confusion matrix for the four AHI scores. The classification accuracy displays the ability of the model to assign a patient to any of the four AHI categories. To gain insight into accuracy of the AHI / RDI prediction disregarding the discrete borders used in categorization, histograms were computed to show the difference between AHI / RDI value scored according to the original labels and the AHI / RDI predicted by our algorithm. Scatter plots visualizing the correlation between the originally score AHI / RDI and the algorithm-predicted AHI / RDI were computed. A robust linear regression model with bi-squared cost function was fitted to the data to compute the correlation between the scored and predicted AHI / RDI [35]. This model was selected to mitigate the effect of outliers. Also, Cohen's kappa values were determined for AHI and RDI prediction.

6.8 Results

6.8.1 Per-event performance

The algorithm was successfully applied on 8803 patient recordings including almost 800k human labeled respiratory events. Applying the rule-based model on the same recordings resulted in the detection of more than 900k respiratory events. This shows that the algorithm exhibits more sensitive scoring behavior compared to the original scorers. When looking at the binary confusion matrix, 74% of all events scored by the experts are also identified by the algorithm while the number of false positive events is about 50% higher than the false negative events, indicating higher sensitivity rather than precision by the algorithm.

Algorithm → Expert ↓	No-event	Event
No-event	8727155 **0.96**	336873 **0.01**
Event	200964 **0.26**	581388 **0.74**

Comparison of the type of events identified by the experts versus the algorithm shows quite some disagreement, as can be observed in the more elaborate multiclass confusion matrix. Much misclassification between events is observed, while the overall specificity is considered high. There must be considered, however, that the specificity is affected by the large class imbalance (regular breathing vs respiratory events).

Algorithm → Expert ↓	No-event	Obstructive apnea	Central apnea	Mixed apnea	Hypopnea	RERA
No-event	8727155 **0.96**	64166 **0.01**	76160 **0.01**	9521 **0.00**	72627 **0.01**	114488 **0.01**
Obstructive apnea	13718 **0.10**	54415 **0.38**	48467 **0.34**	7289 **0.05**	15696 **0.11**	3381 **0.02**
Central apnea	22935 **0.15**	22036 **0.14**	73735 **0.48**	14802 **0.10**	17148 **0.11**	3719 **0.02**
Mixed apnea	318 **0.03**	2766 **0.29**	4292 **0.46**	1266 **0.13**	674 **0.07**	63 **0.02**
Hypopnea	41775 **0.17**	52723 **0.21**	35438 **0.14**	4343 **0.02**	105363 **0.42**	10309 **0.04**
RERA	122372 **0.55**	31679 **0.14**	17677 **0.08**	1798 **0.01**	8409 **0.04**	37830 **0.17**

On the next page notes on the detection capabilities of the different respiratory event types are given. A more detailed discussion is provided in following sections.

Obstructive apnea:

The algorithm agreed on 38% of all obstructive apneas found by the experts. About the same number of obstructive apnea events were classified as central, indicating that our model does correctly identify the 90% flow limitations for apneic events most of the time, but may lack discriminatory capabilities when assessing the associated respiratory effort. From these results can be derived that the algorithm is too sensitive in the identification of a central aspect in respiratory effort and its according thresholds require improvement. For other misclassification by the algorithm resulted in hypopnea classification, implying potential inaccuracies when assessing flow limitations.

Central apnea:

The highest agreement among the event types was found with the detection of central apnea events. About half of all central apneas detected by the experts were also identified by the algorithm. 15% of the central apnea events were missed by the algorithm, expectedly due to disagreement on the duration of such events. During visual assessment, regularly, short apnea events were scored by the experts, which were disregarded by the algorithm.

Mixed apnea:

Detection of mixed apnea events by the algorithm was considered poor, with an overall sensitivity of 13%. Most mixed apneas scored by the experts were identified as either obstructive or central apnea by the model. In clinical setting mixed apneas are reluctantly scored by experts since differentiation from the other types of apnea is not obligatory. Therefore, much label noise considering this type of event was expected, which may explain in part the low agreement between the experts and the algorithm.

Hypopnea:

Hypopnea events were by far the most occurring event type among the respiratory events. An accuracy of 81.5% with similar precision and sensitivity around 45% is considered reasonable performance. Most misclassified events were scored as apnea, indicating disagreement on the peak excursion drop height of the airflow signal.

RERA:

Identification of RERAs was a difficult task for the algorithm. A sensitivity of 17.2% and a F1-score of 19.4% shows extensive considerable disagreement between the model and the experts. The definition given by the AASM is relatively lenient and allows a lot of room for interpretation by the human scorers. Notably, even though most RERA events were missed by the algorithm, a similar number of FP RERAs were found by the algorithm.

An overview of the per-event performance metrics for each type of respiratory event is given in the table below, with all values are in percentages.

	Obstructive apnea	Central apnea	Mixed apnea	Hypopnea	RERA
Accuracy	97.2	97.2	99.5	97.2	96.6
Sensitivity	38.1	47.8	13.5	42.2	17.2
Specificity	98.1	98.0	99.6	98.7	98.5
Precision	23.9	28.8	3.2	47.9	22.3
F1- score	29.4	36.0	5.2	44.8	19.4

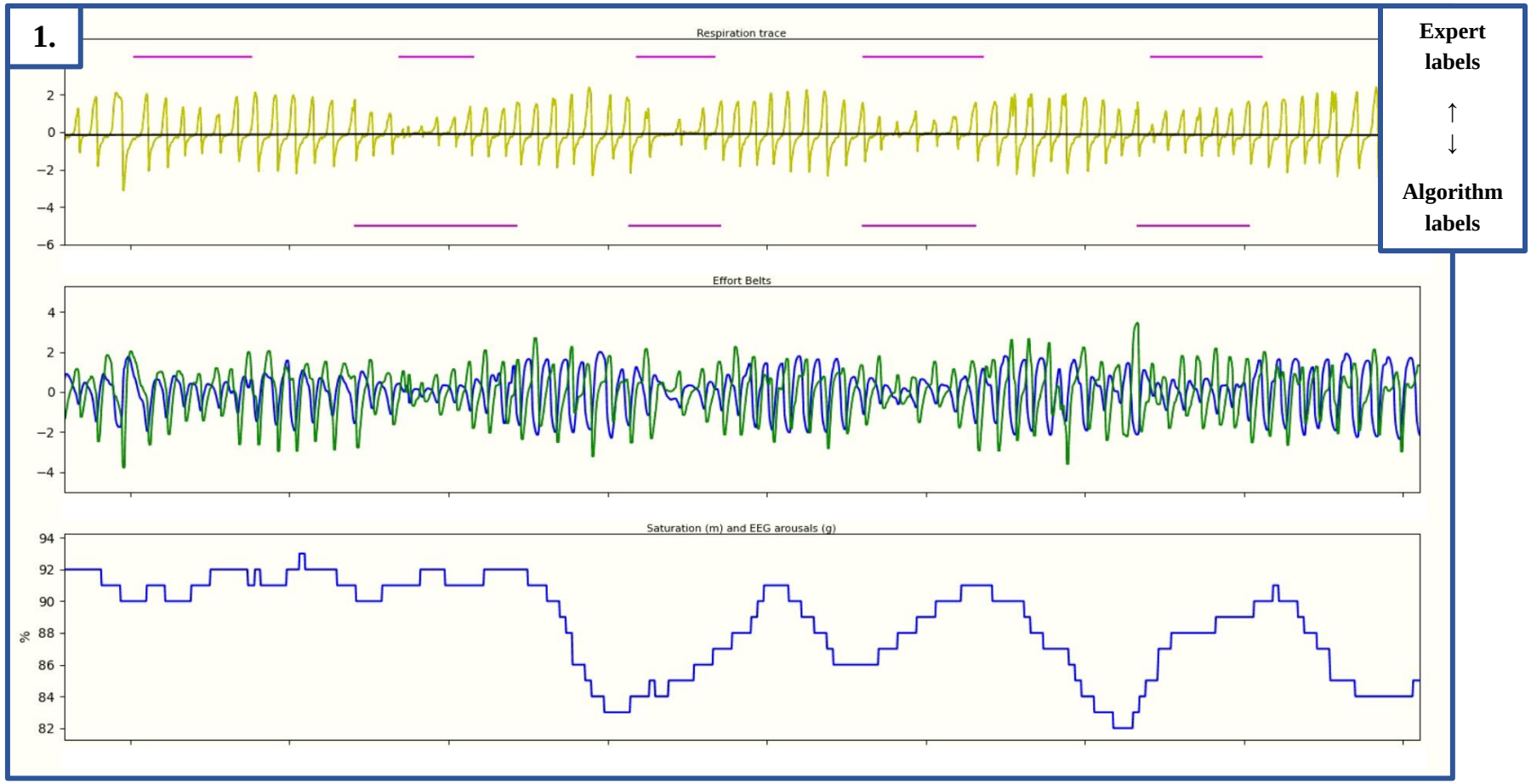
Even though the performance metrics vary greatly between the respiratory event types, misclassification of events often resulted in classification of other event types. Moreover, the binary confusion matrix show that the algorithm and experts agree on almost ¾ of all respiratory events. Grouping the apnea events together and creating a new confusion matrix with 4 classes, i.e., no-event, apnea, hypopnea, and RERA, shows encapsulates the apnea misclassification by the algorithm, see table below. When the algorithm was used to detect apneas, hypopneas and RERAs, an increasing sensitivity is observed corresponding with the severity of the event types. Similar to human scorers, the detection of the prominent apnea events was easier to distinguish from regular breathing than the more obscure RERA events. Besides, disagreement between the apnea and hypopnea type by the human scorers and the algorithm resulted mostly in the detection of the other event type.

Algorithm → Expert ↓	No-event	Apnea	Hypopnea	RERA
No-event	8727155 **0.96**	149847 **0.02**	72627 **0.01**	114488 **0.01**
Apnea	36971 **0.14**	185447 **0.70**	33518 **0.13**	7163 **0.03**
Hypopnea	41775 **0.17**	92504 **0.37**	105363 **0.42**	10309 **0.04**
RERA	122372 **0.55**	51154 **0.24**	8409 **0.04**	37830 **0.17**

	Apnea	Hypopnea	RERA
Accuracy	96.1	97.2	96.6
Sensitivity	70.5	42.2	17.2
Specificity	96.8	98.7	98.5
Precision	38.7	47.9	22.3
F1- score	50.0	44.8	19.4

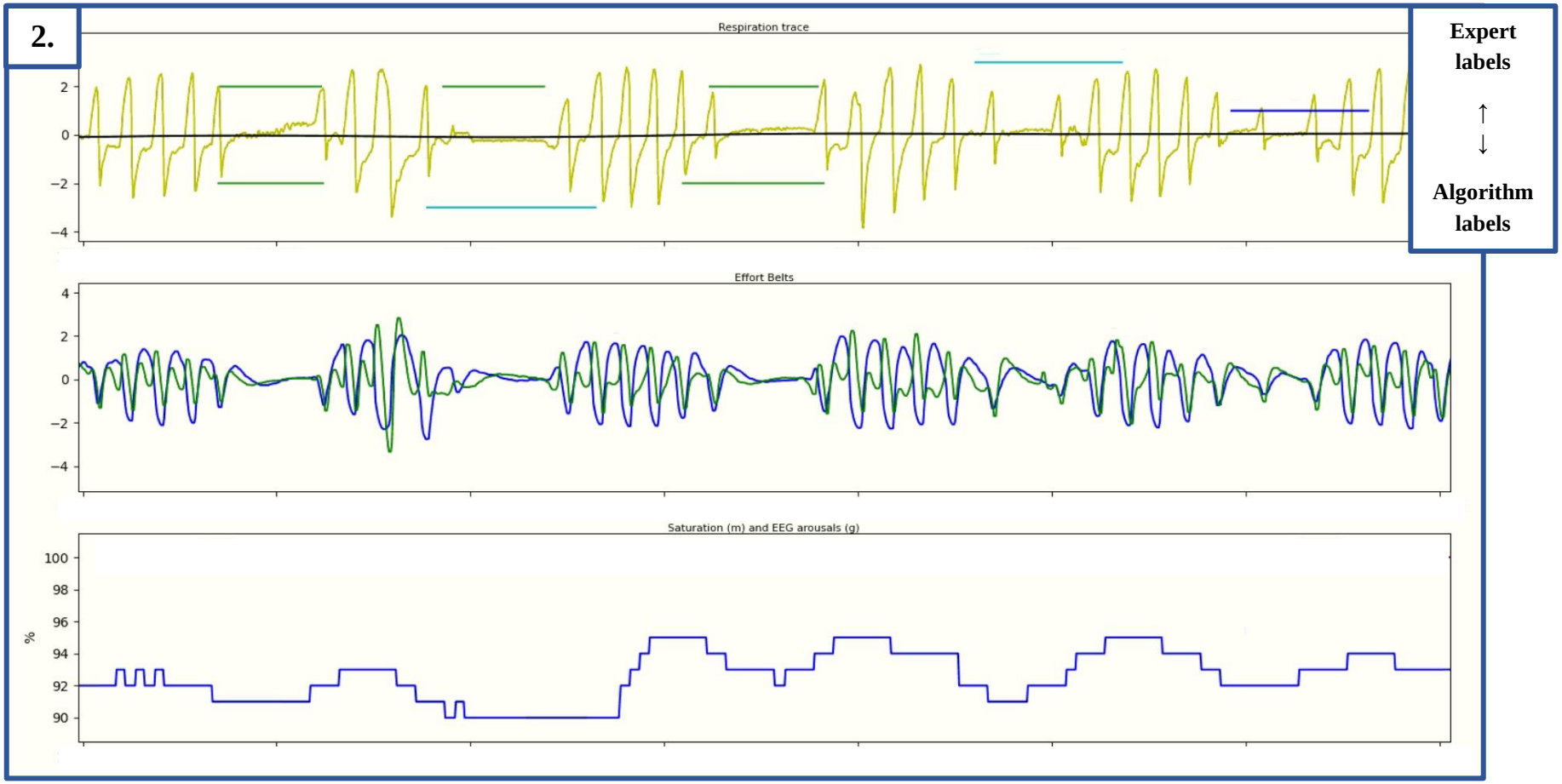
6.8.2 Example segments

On the following nine pages example segments are shown. These segments highlight the scoring behavior of the designed algorithm. For each example, the preprocessed signals are displayed together with the projected original labels and algorithm labels, above and below the airflow signal in the first subplot, respectively. For the second subplot, the blue line represents respiratory effort measured at the abdomen, while the red line reflects respiratory effort determined around the chest. The third subplot shows the oxygen saturation trace.



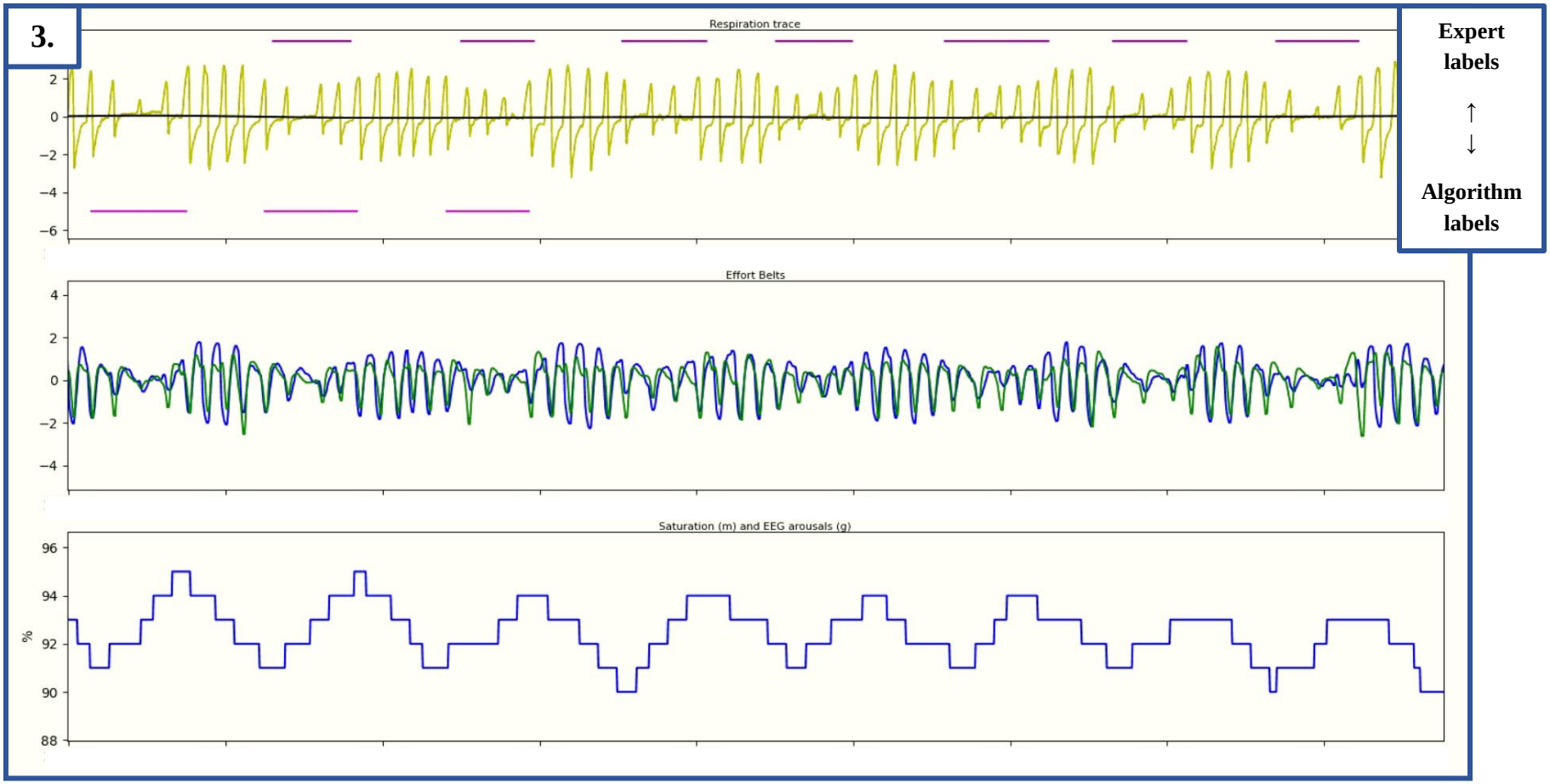
Notes:

This example shows agreement on a chain of hypopnea events by the experts and the algorithm. Flow limitations lay between 30% and 90% and severe saturation drops are present. A leading hypopnea event scored by the experts is disregarded by the model since no matching saturation drop is found. These results highlight the scoring consistency by the algorithm.



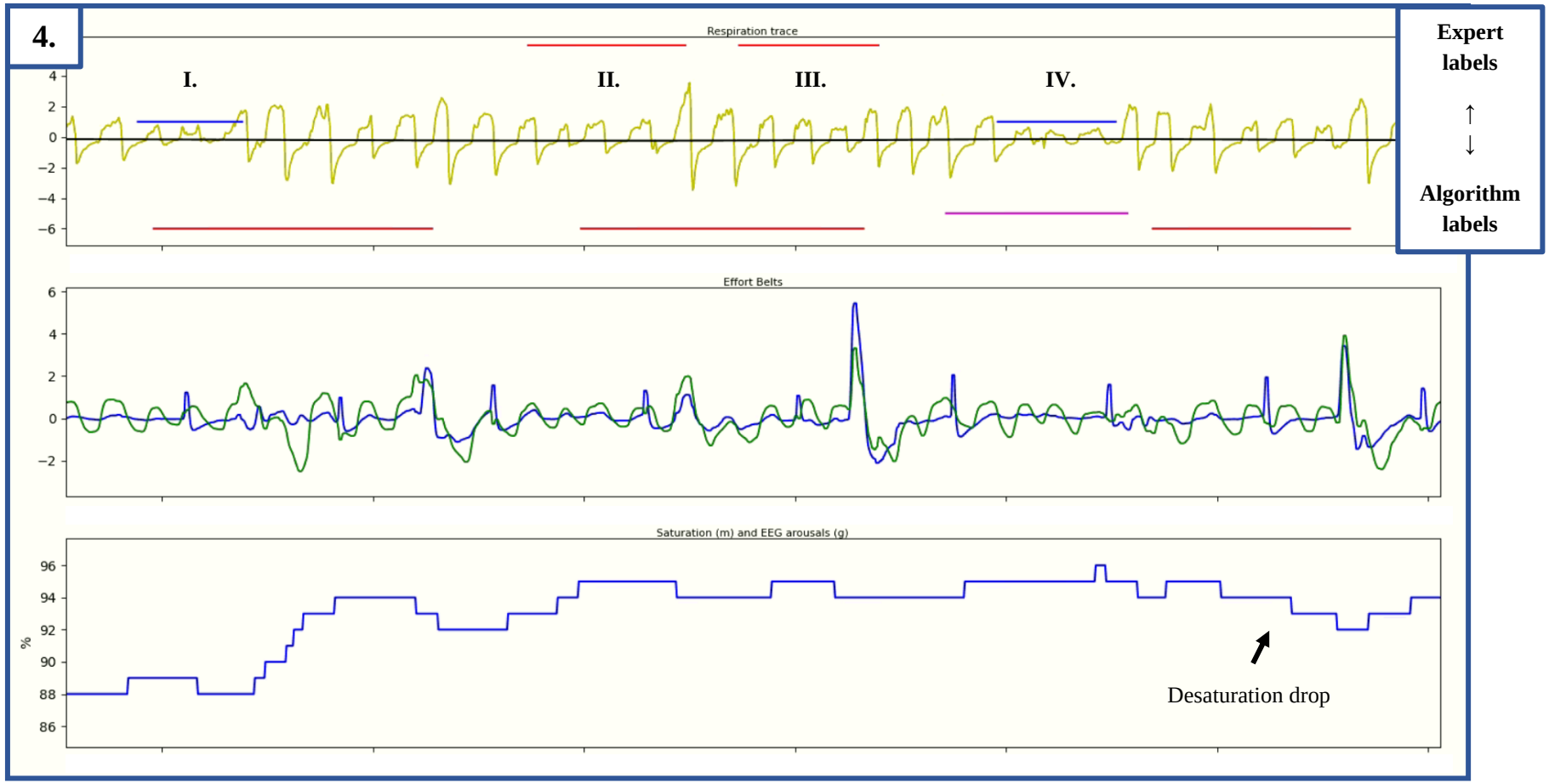
Notes:

This example shows decent overall apnea agreement by the experts and the algorithm. There is concurrence on the first three apneas with one misclassification (central apnea → mixed apnea). Consequently 2 apneas are scored by the experts and not by the model. For both these events the 90% flow limitation with a duration of >10 seconds is not observed by the model and it therefore does not score the last 2 events as apnea. Also, no misclassification as hypopnea occurred since there is no observed 4% desaturation drop.



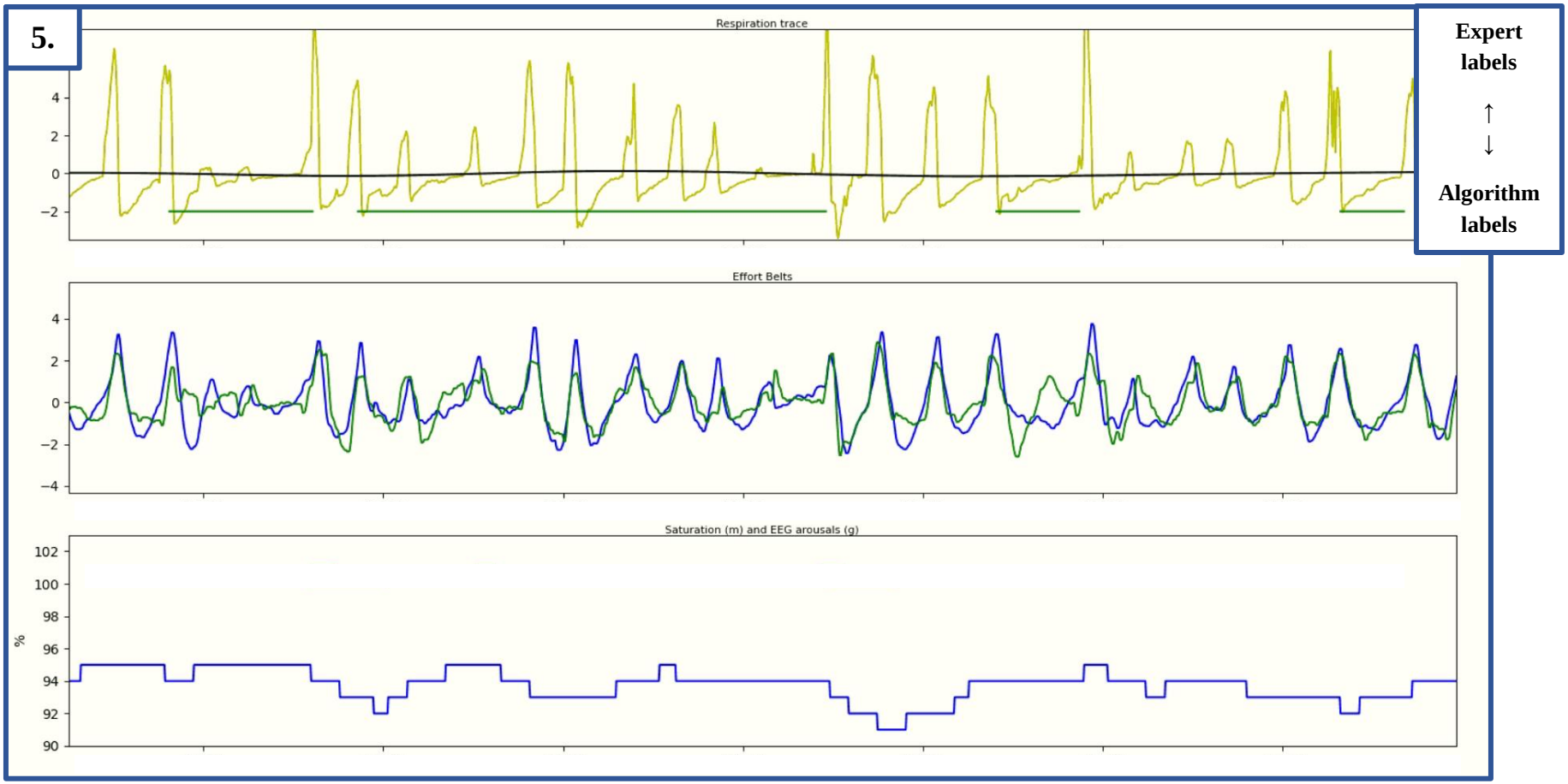
Notes:

This segment shows a chain of 3% and 4% hypopneas scored by the experts. The algorithm, however, disregards the 3% hypopneas as instructed. A leading hypopnea was missed by the experts for unclear reasons. Although the experts are instructed to follow the 4% rule when scoring hypopneas, inconsistency is observed regularly. Especially when severe flow limitations of near apneic criteria are observed (e.g., 60-90% drop from baseline peak excursion).



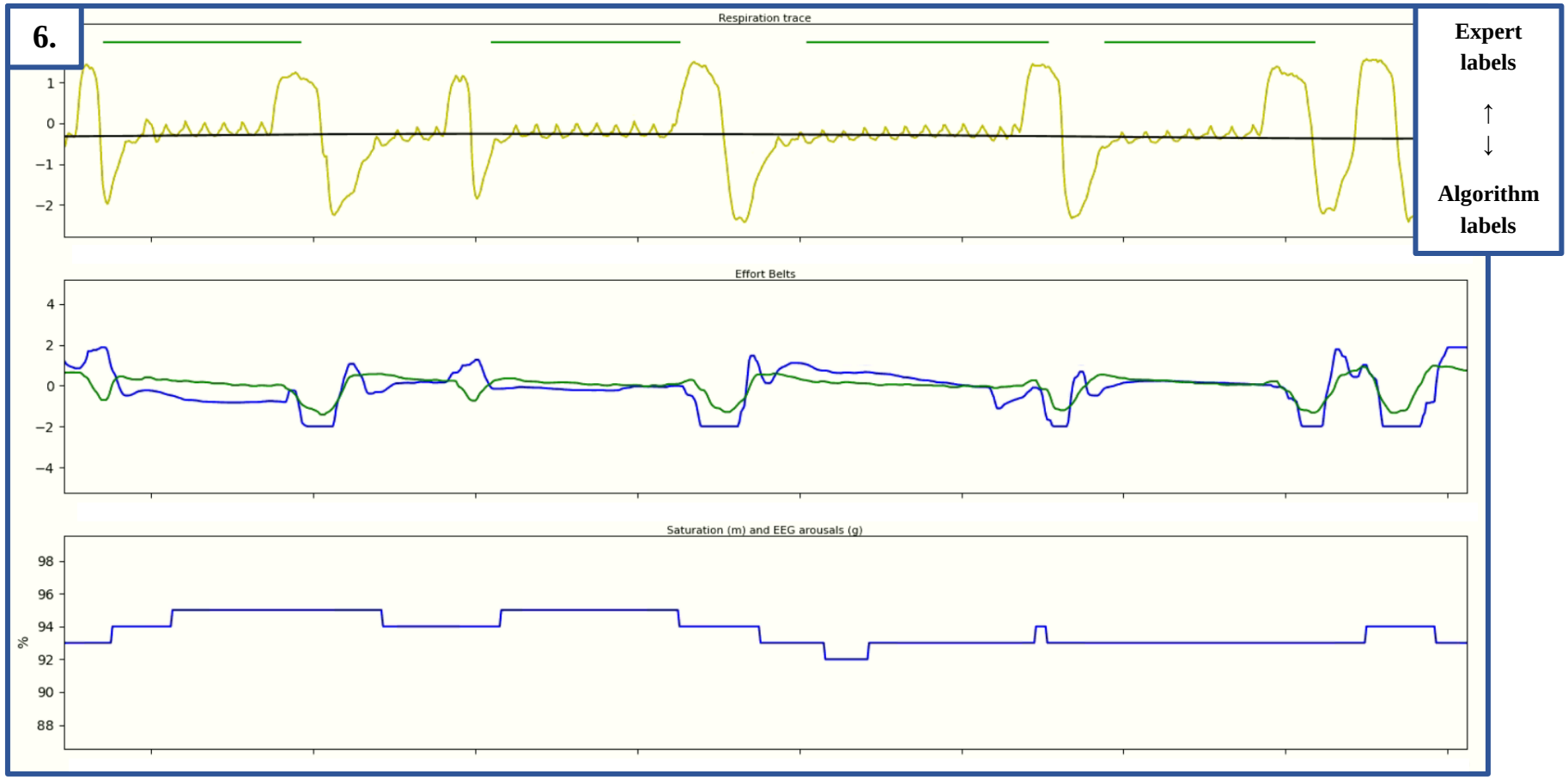
Notes:

Both the experts and the algorithm agree that multiple respiratory events are present in this segment, yet there is disagreement on the location and type. The algorithm does not agree with the 90% airflow peak excursion for events (**I.** & **IV.**). Instead, for event **I.**, a RERA is detected by the algorithm while for event **IV.** a hypopnea is found matching the following 4% desaturation drop. For the 2 RERAs scored by the experts (events **II.** and **III.**) the model only identifies a single large RERA indicating that our model might have difficulties accurately determining start and end locations of events.



Notes:

This example segment shows that very irregular breathing with many spikes may lead to false positive predictions. The model did not accurately determine a ventilatory flow envelope, as high spikes increase baseline excursion. Very large apneas may be the result, and perhaps a good indicator for this problem.



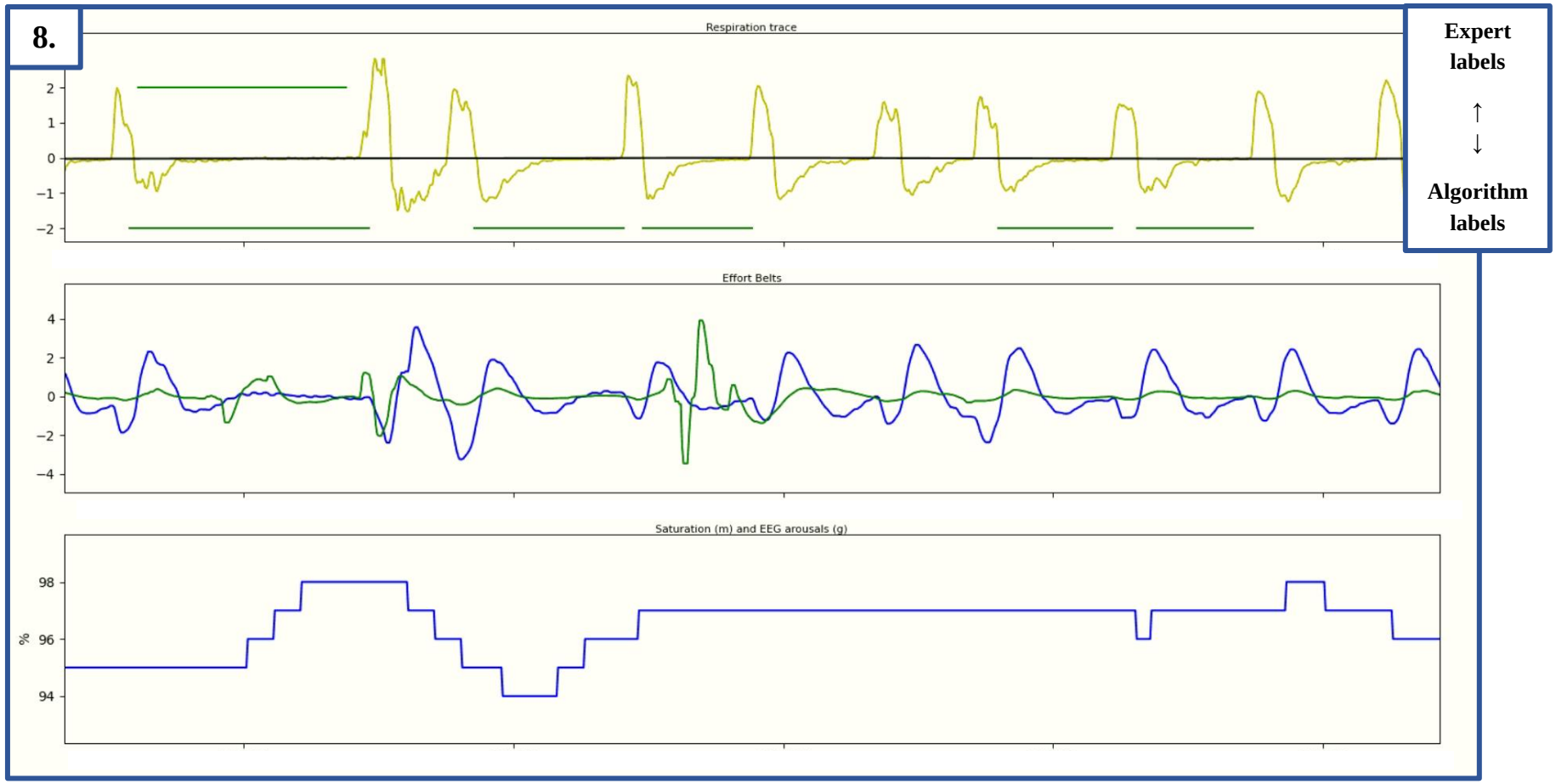
Notes:

This segment shows a chain of missed central apneas by the algorithm. Likely due to the high AHI of this patient, the respiratory flow envelope was inaccurately determined, leading to inaccurate determined baseline peak excursion (lower). For this reason, the algorithm did not identify the flow limitations to exceed the 90% excursion drop rule, hence no prediction of apnea events. Since no 4% desaturation drops were detected, also no hypopneas were identified by the model.

7.

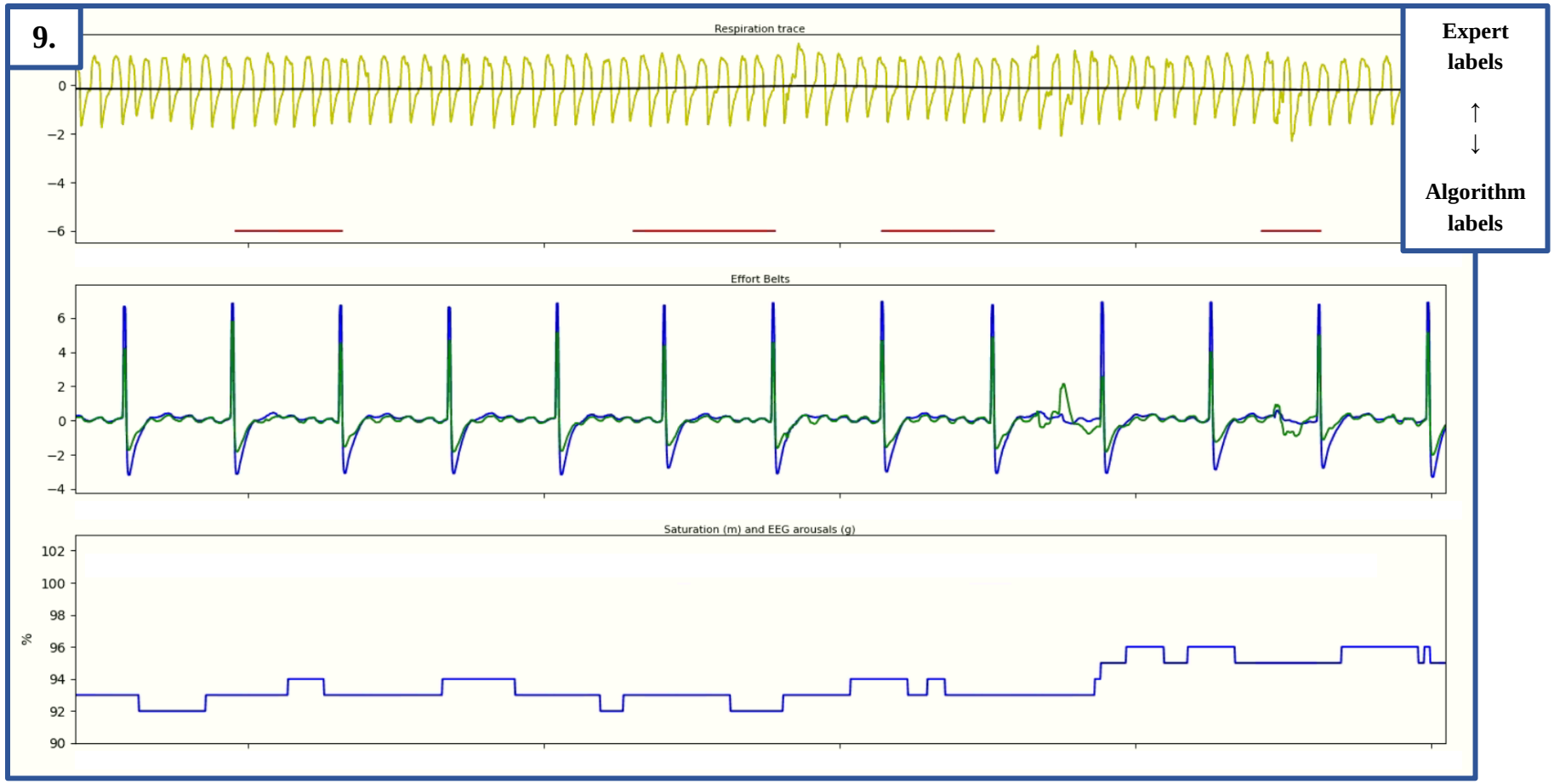
Notes:

This example shows that highly irregular breathing may lead to very large apneas detected by the model. The duration of these apneas is not physiologically feasible and need to be addressed in later model improvement. An easy solution would be removing such long events. Better would be reassessing such areas and identifying possible shorted events within that area.



Notes:

Low respiratory rates (e.g., bradypnea) display similar characteristics as flow limitations (see Section 0), which resulted in false positive respiratory events.

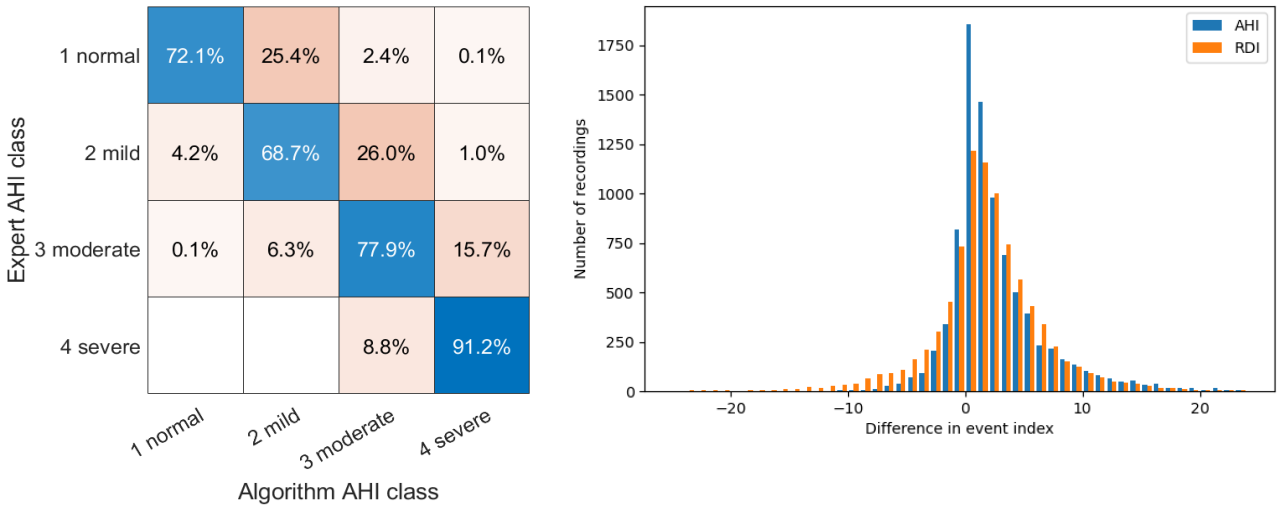


Notes:

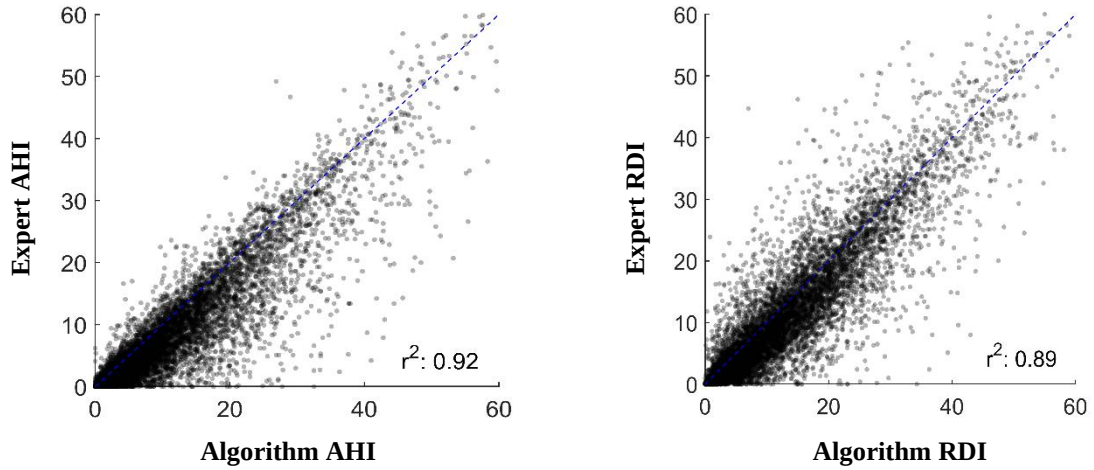
This example shows regular breathing (according to the experts), whereas the algorithm detected numerous RERAS. Respiratory effort shows spikes with a stable frequency around 0.08 Hz caused the model to detect EEG arousals, which lead to FP RERA events. This emphasizes a limitation of automated scoring by algorithms. Automated models are prone to artifacts in data which may lead to false results. Experienced human scorers easily recognize such artifacts and will disregard such spikes during scoring.

6.8.3 Global performance

By grouping together all respiratory events detected in a patient the AHI and RDI values were computed using the equations from Section 6.7.2. For all patients AHI categorization by the experts and the algorithm was compared, resulting in the AHI confusion matrix below. The original labels and the algorithm agreed on apnea severity for 72% out of all patients resulting in an according Cohen's kappa value of 0.59. It is notable that most misclassifications of the algorithm resulted in false positives into the neighboring AHI categories. This effect is best visualized in the histograms where the difference between the experts and the algorithm for both event indexes can be observed. The unimodal and symmetrical shape shows a decrease in number of false positives as the difference between the AHI / RDI scored by the experts and the algorithm increased. With respect to the 17% overall RERA sensitivity by the algorithm, the global RDI performance seems relatively accurate. The bell curved shape of the RDI histogram indicates that the large numbers of FP and FN RERAs, as seen the in per-event confusion matrices, are quite evenly distributed among all patients, with and without a large overall RDI. Therefore, while the per-event performance of RERA detection is quite poor, prediction of the RDI of patients is significantly better.



The AHI and RDI correlation can be nicely displayed in scatterplots. An according r^2 value of 0.92 for the AHI and r^2 value of 0.89 for the RDI was found. Also here, tendency to overpredict respiratory events by the algorithm is observed. Besides, the RDI scatter plots shows a larger variability than the AHI scatter plot, implying an increased detection error when scoring by the algorithm includes the detection of RERA events.



6.9 Discussion

This chapter described how rule-based algorithms can be used to automatically detect respiratory events during sleep. The implementation of hyperparameters that incorporate the important ambiguity within the rules of the AASM scoring manual, allows manipulation and hyperparameter tuning to mimic the scoring behavior exhibited by human annotators. Evaluation on event-per-event granularity and global patient level was performed. The per-event performance varied much between the different types of respiratory events and the overall detection efficiency increased with the severity of the respiratory event types, demonstrating that more harmful events in fact more clearly display discernible physiological change that is detectable by automated models. 74% of all events identified by experts were also found by the proposed model, which is comparable to current state-of-the-art models (see Section 3.2.1), but among these events much misclassification of other event types occurred. Accurate discrimination between the individual event types holds important clinical value. For instance, the type of breathing assistance and overall apnea treatment may vary for different underlying pathology leading to apnea. Specifying the type of apnea will therefore provide aid in improving personalized patient care. The ability to discriminate various respiratory events appears promising using a rule-based algorithm. Work by *ElMoaqet et al (2020)* is the only study in literature that shows that discrimination between apnea types (not hypopnea or RERA) is possible with a respective accuracy, sensitivity, specificity, precision, and F1-score of 83.4%, 88.5%, 82.5%, 46.6%, and 42.7%. Comparison with the rule-based model shows very similar performance metrics with an accuracy, sensitivity, specificity, precision and F1-score of 96.1%, 70.5%, 96.8%, 38.7%, and 50.0%, respectively. This shows that the proposed model can already identify respiratory events on event-level granularity with equivalent efficiency as the current leading models.

The high global performance indicates that AHI prediction based on the specific respiratory events is feasible regardless of reduced discriminatory accuracy by the rule-based model. An r^2 value of 0.92 for the AHI and 0.89 for the RDI shows a strong correlation between the algorithm and human

scorers and the AHI confusion matrix displays an overall agreement range of 70-90%. A Cohen's kappa value of 0.59 was determined. Global assessment of patients appears an easier automation task compared to per-event granularity assessment and the proposed rule-based model can compete with state-of-the-art models found in literature [29]–[35].

The 50% rule to assess the correctly classified respiratory events by the algorithm with respect to the original labels was an arbitrary decision and might be strict considering the limited clinical relevance to any exact start and end location of an event. Small events that are slightly shifted might be unfairly penalized by our assessment method. In addition, the duration of expert annotations is questionable. "Event hot-keying", using a predefined event duration rounded to tens of seconds, is a common strategy to accelerate manual scoring. Events with a duration just below the 10 second minimum but also longer events are prone to incorrect labelling by human scorers, possibly decreasing the performance metrics determined for the algorithm.

Since the proposed algorithm is a first attempt in building a fully rule-based model to score and differentiate between respiratory event types, the performance is considered highly promising. This prototype model is likely to improve in accuracy with future optimization iterations. Experts in the field of sleep medicine from various leading medical institutions are involved in the assessment during such optimization iterations and future work will focus on the following limitations.

Main limitations of this work come forth out of the arbitrary decision making during the algorithm design. As discussed in chapter 0, the AASM manual is a textual description of the respiratory events and stringent specifics are absent, purposely to prevent excessive demarcation. Yet, ambiguities become accentuated when trying to comply to these rules in an ultimately literal language, Boolean logic. The proposed algorithm will never consider clinical information outside of the "learned" AASM criteria or exhibit patient-dependent behavior during scoring, like human scorers. Strict compliance by any rule-based algorithm for automated scoring holds advantages when it comes to robustness, but its limitation shows when dynamic interpretation is required.

Additional in-depth hyperparameter-tuning will help identify the important event characteristics that human scorers use to differentiate between event types. Tweaking the compartmentalized modules of the algorithm might even hint to specific signal features that are not described by the AASM criteria yet can prove to be excellent hallmarks in the detection of events, e.g., recovery breaths.

The more extreme errors by the algorithm as seen in example segments 5, 6, and 7, indicate that the computation of the airflow envelope can be improved. Patients with a large RDI often display very irregular breathing, which causes inaccurate results when using the current method to compute a baseline peak excursion. Even though such errors do not occur in most patients, improvement is needed to enhance the robustness and reliability when applying this model in

clinical practice. Future development will include revision of the envelope computation method. Introduction of a negative baseline peak excursion might be useful, which can serve as an additional information feature when assessing peak excursion drops. Segments where inspiration and expiration display asymmetrical amplitudes are examples of interest in which a combination of positive and negative peak excursion might yield better performance than solely focusing on inspiratory signal features.

Effort signals appear to be highly irregular and vary substantially among, but also within, patient recordings. Large and sudden amplitude changes outside of realistic underlying physiological representation, likely due to movement artifacts, constrain detection of actual physiological change caused by respiratory events. RIP signals are prone to artifacts by movement due to its measurement location and high measurement sensitivity [48]. Assessing the quality of the individual effort signals measured at the abdomen and chest before combining both might improve better results when discriminating between obstructive and central apnea. Besides, thoracoabdominal asynchrony reflected in the effort signals is a hallmark for obstructive apnea. Like human scorers, the algorithm may be able to identify this hallmark which may in turn lead to improved discrimination between apnea type.

In manual analysis, experts learn to implicitly visually discount artifacts. As seen in example segment 9, artifacts easily lead to erroneous results by an automated rule-based model. More elaborate preprocessing may reduce the probability of present artifacts in data.

Even though the preprocessing applications were limited, down sampling of the data and the use of filters may still have impacted the signal quality and thus affect the model performance. The applied preprocessing steps in this work were expected to vastly reduce artifacts while having a limited negative impact on the quality of the included signal and their reflected physiological information.

6.10 Conclusion

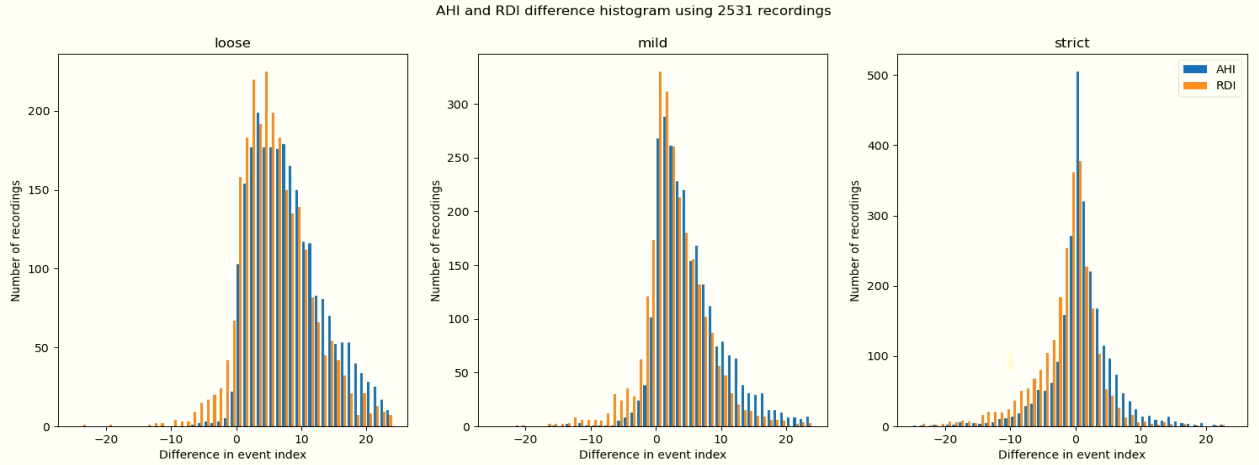
This work proposed an innovative and completely original modelling approach to automatically score respiratory events during sleep, which unlike deep learning approaches is constrained by human knowledge. Global patient assessment by the model resulted in strong agreement with the current gold standard, manual scoring. While per-event scoring performance showed comparable to current state-of-the-art models, further development is required before clinical implementation is feasible. Building upon the current results, iterations of experimental assessment of the model compartments and the according hyperparameters is likely to increase current performance metrics up to human level scoring. Using the AASM criteria as a blueprint to design a rule-based model architecture is a promising supervised automation method to capture and imitate human scoring behavior within a data-driven framework.

7 Inter-rater agreement experiment

The guidelines for scoring respiratory events manually have evolved over the years but remained largely driven by consensus. The AASM therefore updates regularly as consensus by experts changes over time. This dynamic aspect of the gold standard makes it difficult to assess automated scoring models. Not only since often older data is used to validate new models, but also because multi-scorer data is rare. Manual scoring of a single recording is tedious and requires up to multiple hours and requesting various human scorers to assess the same recording is practically challenging. Instead of rescoring full patient recordings, interesting segments can be cut from original records and provided to different assessors for scoring. AASM accredited sleep centers, such as the MGH, have stringent ongoing requirements for documenting and maintaining inter-scorer reliability of over 85%. Monthly, sleep technicians are required to label numerous event segments, which thereafter get assessed by a board of experts who are responsible for governing the inter-rater variability among all associated local scorers. In this way, inter-rater agreement can thus be assessed without the need of scoring complete recordings by all AASM accredited technicians. A limitation of this method, however, is that it is not possible to compute AHI related performance metrics based on recording segments since these should be based on full recordings. To reconstruct this assessment method by the AASM I created a tool (with help from MGH colleagues) to rescore segments that could also be evaluated by our rule-based model.

7.1 Sample selection

Multiple experts were asked to use the tool and rescore 1020 3-minute sample segments. This experiment ideally includes segments containing each of the different respiratory event types, encompassing samples that are relatively simple and relatively difficult to identify. By manipulation of the hyperparameters of our rule-based algorithm I could alter the model to exhibit more, or less, sensitive scoring behavior, i.e., by increasing and decreasing the thresholds within the compartmentalized assessment modules. For instance, decreasing computed airflow envelope thresholds lead to more sensitive flow limitation detection, which in turn leads to increased apnea and hypopnea detection. This effect was nicely captured when looking at the histograms showing the respiratory index difference between the model and the original labels for three different settings: loose, mild, and strict. The left histogram shows how “loose” hyperparameter settings lead to more respiratory event detections reflected in an increased AHI and RDI with respect to the original labels. This resulted in increased overprediction by the model. Conversely, the “strict” hyperparameter settings resulted in a clear peak around 0, indicating that for most patients the AHI / RDI computed by the algorithm was similar as the AHI / RDI determined by the human scorers.



All events from 600 patients were scored by the algorithm using all three hyperparameter settings. When all three models agreed on a specific event, this event was tagged with 'high-conviction', while events that were only found by the loose model were accredited with the tag "low-conviction". Next, all events were compared with the original labels, from which TP, FP, and FN events could be determined. An even number of high- and low-conviction TP, FP, and FN events were sampled among the different event types (including obstructive apnea, central apnea, mixed apnea, 3% hypopnea, 4% hypopnea, and RERA). This heterogenous group of event segments with all including preprocessed (using similar steps explained in the previous chapter) respiration and EEG signals were extracted from original recordings using customized algorithms. Besides the three AASM required signals, we included all signals human scorers may use to score respiratory events in their conventional setting.

7.2 Experiment setup

All 1020 segments extracted from 600 patients were presented to the rescoring experts. The guidelines presented on the next two pages were sent to all included scorers. This rescoring experiment is ongoing, and preliminary results are currently coming in. Once we obtain results from approximately 5 scorers, we will use a majority vote system to label each segment. Creating new labels using the combined expertise of multiple scorers is expected to result in labels of better quality than the current 1-scorer labels.

Respiratory event rescoring tool guidelines

AIM:

Currently, we have constructed a prototype computer model that detects the conventional respiratory events scored in clinical PSG's according to the AASM. With this model we can automatically identify and discriminate obstructive, central, and mixed apneas, hypopneas (3% + arousal and 4% desaturation) and respiratory effort related arousals. The next step consists of validating and optimizing the sub-algorithms. We have built a tool that allows for easy rescoring of ±1000 respiratory events. With this tool we aim to:

1. *Gain more insight in the scoring behavior of sleep experts. By analyzing the results from experts from various institutions we can identify the important scoring characteristics on event-level and quantify the inter-rater variability.*
2. *Perform parameter optimalization of our algorithm. Using new labels derived from multiple experts we can further improve our model using labels that are more robust and of better quality than current 1-scorer data.*

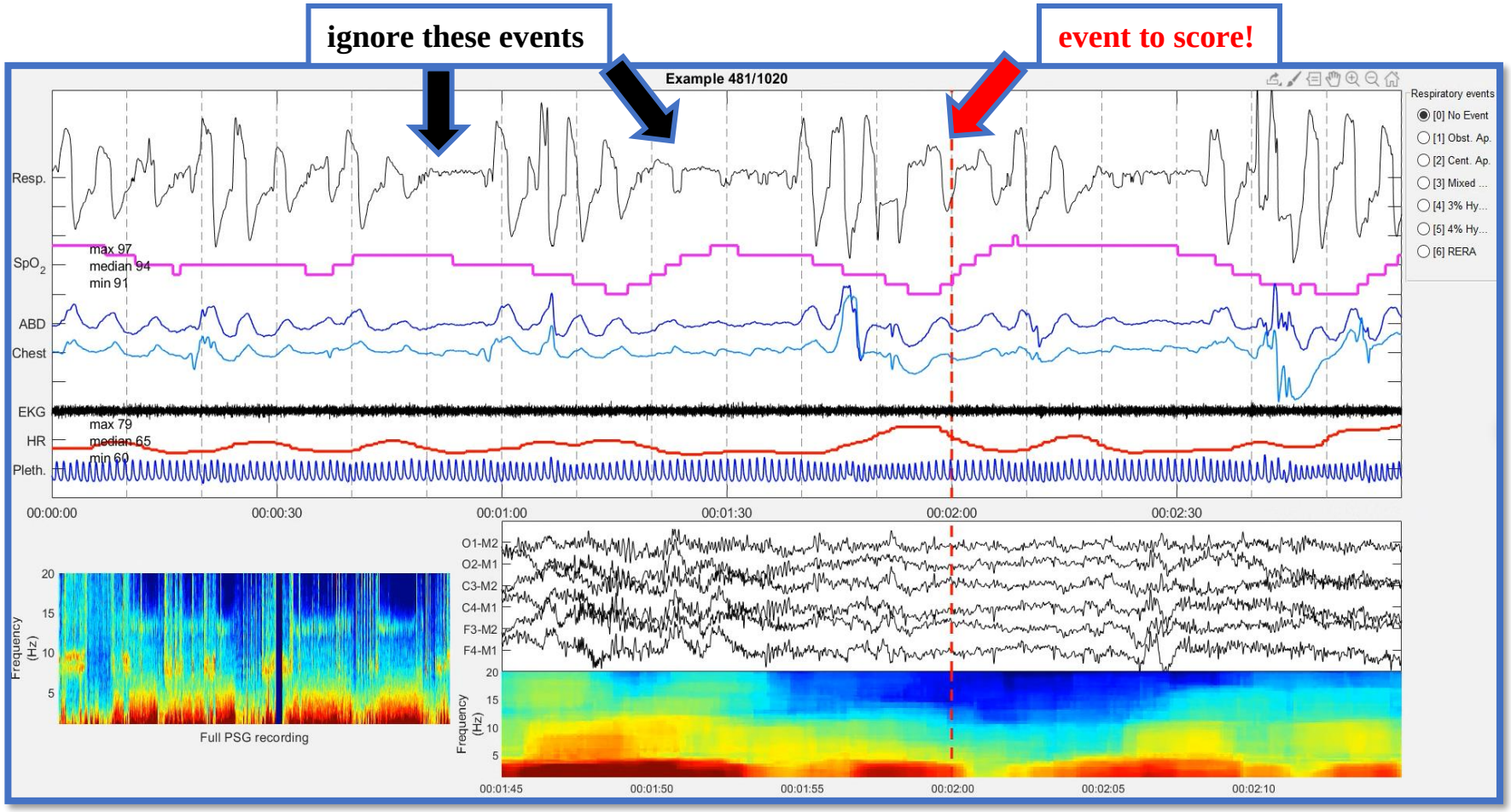
Instructions:

The goal of this experiment is to assess and score each example by selecting one of the following options: 'obstructive apnea', 'central apnea', 'mixed apnea', '3% hypopnea', '4% hypopnea', 'RERA', or 'No event'. This is how to interpret each example:

- ***The location of the possible event is specified by the red-dotted line***
- ***Each event shows 2 minutes of leading context, and 1 minute of trailing context, except for the EEG traces that show 15 seconds of both leading and trailing context.***

Note: Only if and when an event overlaps the location of the red-dotted line, the event should be scored accordingly. Other present events withing the example window should be ignored!

Besides the EEG traces for arousal detection, only ventilation, saturation, and respiratory effort signals are required for scoring according to the AASM. However, to aid the participants of this experiment ECG, heart rate, oxygen plethysmography and EEG spectrogram signals are provided for each example. On the next page an example of the tool layout can be found.



Keyboard functions:

The tool is designed for simplicity and efficiency use. Therefore, we assigned the following keyboard keys to use while scoring.

Key:	*Function:*
number 0	*No Event*
number 1	*Obstructive Apnea*
number 2	*Central Apnea*
number 3	*Mixed Apnea*
number 4	*3% Hypopnea (3% desaturation drop and/or arousal)*
number 5	*4% Hypopnea (4% desaturation drop)*
number 6	*Respiratory effort related arousal*

Key:	Function:
arrow →	next example
arrow ←	previous example
arrow ↑	scale up EEG
arrow ↓	scale down EEG

7.3 Preliminary results

With the results from the first two scorers that completed the experiment confusion matrices and Cohen's kappa values were determined to assess the level of agreement. The binary confusion matrix shows that expert 2 identified 96% of all event segments determined by expert 1. Grouping all detected apneas together resulted in a 95% agreement by both experts, while this was true for 77% of all hypopneas. Nearly all disagreement of apnea and hypopnea segments resulted in detection of the other type of respiratory event. The full multi-class confusion matrix indicates that misclassification significantly increases, and the overall event agreement declines when experts try to differentiate between all included respiratory events. A Cohen's kappa value of 0.43 was determined. The below results demonstrate that considerable inter-rater variability is observed when two experts score the exact same PSG segments on event-level granularity. Global assessment by AHI and RDI computations is unlikely to show a large variability among the human scorers, since a high agreement is observed when events are grouped together.

Scorer 1 → Scorer 2 ↓	No-event	Obstructive apnea	Central apnea	Mixed apnea	Hypopnea (3%)	Hypopnea (4%)	RERA
No-event	0.46	0.05	0.11	0.01	0.16	0.07	0.14
Obstructive apnea	0.01	0.61	0.32	0.00	0.00	0.06	0.00
Central apnea	0.00	0.01	0.91	0.04	0.03	0.00	0.00
Mixed apnea	0.03	0.14	0.48	0.34	0.00	0.00	0.00
Hypopnea (3%)	0.07	0.08	0.05	0.00	0.65	0.13	0.01
Hypopnea (4%)	0.01	0.11	0.09	0.02	0.11	0.66	0.00
RERA	0.30	0.03	0.03	0.00	0.32	0.11	0.22

Scorer 1 → Scorer 2 ↓	No-event	Apnea	Hypopnea	RERA
No-event	0.46	0.18	0.23	0.14
Apnea	0.01	0.95	0.04	0.00
Hypopnea	0.03	0.20	0.77	0.00
RERA	0.30	0.05	0.43	0.22

Scorer 1 → Scorer 2 ↓	No-event	Event
No-event	0.46	0.54
Event	0.04	0.96

Binary: **Cohen's kappa: 0.40**
semi-Multi: **Cohen's kappa: 0.47**
Multiclass: **Cohen's kappa: 0.42**

Algorithm → Scorer 1 ↓	No-event	Obstructive apnea	Central apnea	Mixed apnea	Hypopnea (3%)	Hypopnea (4%)	RERA
No-event	0.58	0.04	0.04	0.03	0.09	0.05	0.17
Obstructive apnea	0.14	0.52	0.13	0.05	0.05	0.11	0.00
Central apnea	0.22	0.17	0.25	0.19	0.05	0.08	0.04
Mixed apnea	0.09	0.13	0.43	0.30	0.00	0.00	0.04
Hypopnea (3%)	0.36	0.11	0.04	0.01	0.30	0.08	0.11
Hypopnea (4%)	0.28	0.15	0.03	0.00	0.10	0.38	0.06
RERA	0.48	0.06	0.00	0.01	0.24	0.01	0.20

Algorithm → Scorer 1 ↓	No-event	Apnea	Hypopnea	RERA
No-event	0.58	0.10	0.14	0.17
Apnea	0.19	0.66	0.13	0.02
Hypopnea	0.32	0.17	0.43	0.08
RERA	0.48	0.07	0.26	0.20

Algorithm → Scorer 1 ↓	No-event	Event
No-event	0.58	0.42
Event	0.28	0.72

Binary: **Cohen's kappa: 0.27**
semi-Multi: **Cohen's kappa: 0.34**
Multiclass: **Cohen's kappa: 0.26**

The overall event agreement of 96% as observed between the two human scorers decreased to 72% and 79%when comparing the results of the algorithm with scorer 1 and 2, respectively. Apnea and hypopnea agreement among the included event segments reduced to 66% and 43% with scorer 1 and 71% and 45% with scorer 2. Increased misclassification between the no-event and apnea/hypopnea class was observed. RERA detection between experts and the algorithm appeared similarly poor. Cohen's kappa values within the range of 0.25-0.35 were determined. These results suggest that the current algorithm does show a clear agreement with human annotators when scoring respiratory events. Human-level scoring performance, however, is not yet achieved and requires further experimentation.

Algorithm → Scorer 2 ↓	No-event	Obstructive apnea	Central apnea	Mixed apnea	Hypopnea (3%)	Hypopnea (4%)	RERA
No-event	0.49	0.10	0.05	0.04	0.13	0.06	0.13
Obstructive apnea	0.13	0.45	0.20	0.05	0.06	0.11	0.01
Central apnea	0.19	0.14	0.33	0.19	0.09	0.06	0.01
Mixed apnea	0.07	0.24	0.31	0.34	0.00	0.00	0.03
Hypopnea (3%)	0.27	0.21	0.07	0.00	0.35	0.08	0.03
Hypopnea (4%)	0.23	0.14	0.07	0.04	0.09	0.37	0.07
RERA	0.41	0.08	0.00	0.03	0.14	0.03	0.32

Algorithm → Scorer 2 ↓	No-event	Apnea	Hypopnea	RERA
No-event	0.49	0.19	0.19	0.13
Apnea	0.14	0.71	0.13	0.02
Hypopnea	0.24	0.25	0.45	0.05
RERA	0.41	0.11	0.16	0.32

Algorithm → Scorer 2 ↓	No-event	Event
No-event	0.49	0.51
Event	0.21	0.79

Binary: **Cohen's kappa: 0.27**
semi-Multi: **Cohen's kappa: 0.30**
Multiclass: **Cohen's kappa: 0.25**

In Section 6.9 I discussed the unimportance of a precise start and end location of a respiratory event. Both visually, but also computationally, it is hard to determine exactly when a respiratory event initiates, which does not make much difference when identifying severe flow limitations, but when computing the duration of an event, it becomes crucial. Especially short events that barely exceed the minimum 10 second duration threshold may result in a significant number of false positives and false negatives by scorers. Since it is difficult to visually assess the exact duration of an event, human scorers are expected to regularly include shortened events. To test this hypothesis, I increased the algorithm sensitivity for the detection of flow limitations by decreasing the duration threshold to 8 seconds. Expectedly this leads to increased event detection combined with reduced specificity by the algorithm. On the following pages new confusion matrices are shown for the more sensitive algorithm with respect to the same 2 human scorers.

Algorithm → Scorer 1 ↓	No-event	Obstructive apnea	Central apnea	Mixed apnea	Hypopnea (3%)	Hypopnea (4%)	RERA
No-event	0.43	0.11	0.06	0.02	0.20	0.07	0.11
Obstructive apnea	0.03	0.64	0.20	0.09	0.03	0.02	0.00
Central apnea	0.07	0.26	0.33	0.22	0.03	0.06	0.02
Mixed apnea	0.00	0.22	0.57	0.22	0.00	0.00	0.00
Hypopnea (3%)	0.20	0.20	0.04	0.03	0.40	0.10	0.05
Hypopnea (4%)	0.09	0.23	0.06	0.01	0.09	0.50	0.03
RERA	0.33	0.17	0.02	0.02	0.34	0.02	0.09

Algorithm → Scorer 1↓	No-event	Apnea	Hypopnea	RERA
No-event	0.43	0.19	0.27	0.11
Apnea	0.05	0.86	0.07	0.01
Hypopnea	0.14	0.27	0.54	0.04
RERA	0.33	0.22	0.36	0.09

Algorithm → Scorer 1 ↓	No-event	Event
No-event	0.43	0.57
Event	0.12	0.88

Binary: **Cohen's kappa: 0.33**
semi-Multi: **Cohen's kappa: 0.40**
Multiclass: **Cohen's kappa: 0.29**

A rather significant increase in sensitivity was observed when decreasing the minimum duration for event detections. Now, instead of 72% and 79%, the algorithm detected 93% and 88% of the events identified by the experts. 90% and 56% of all apneas and hypopneas found by expert 1 were identified by the modified algorithm, compared to 66% and 43% agreement of the original model with expert 1. For expert 2, the apnea and hypopnea agreement increased from 71% and 45% to 86% and 54%, for apneas and hypopneas, respectively. Apart from mixed apneas and RERAs, the number of agreed upon events between the algorithm and the experts increased. The decreased number of congruent no-event segments indicates a decreased specificity by the model.

Algorithm → Scorer 2 ↓	No-event	Obstructive apnea	Central apnea	Mixed apnea	Hypopnea (3%)	Hypopnea (4%)	RERA
No-event	0.33	0.19	0.10	0.04	0.20	0.07	0.07
Obstructive apnea	0.03	0.48	0.29	0.11	0.02	0.05	0.01
Central apnea	0.06	0.24	0.36	0.29	0.01	0.03	0.01
Mixed apnea	0.00	0.31	0.41	0.24	0.00	0.00	0.03
Hypopnea (3%)	0.07	0.29	0.05	0.01	0.47	0.08	0.03
Hypopnea (4%)	0.06	0.23	0.08	0.04	0.10	0.46	0.01
RERA	0.27	0.16	0.00	0.03	0.27	0.05	0.22

Algorithm → Scorer 2 ↓	No-event	Apnea	Hypopnea	RERA
No-event	0.33	0.33	0.27	0.07
Apnea	0.04	0.90	0.05	0.02
Hypopnea	0.06	0.36	0.56	0.02
RERA	0.27	0.19	0.32	0.22

Algorithm → Scorer 2 ↓	No-event	Event
No-event	0.33	0.67
Event	0.07	0.93

Binary: **Cohen's kappa: 0.23**
semi-Multi: **Cohen's kappa: 0.30**
Multiclass: **Cohen's kappa: 0.25**

There must be considered that these results are preliminary and conclusive statements are yet to be determined. The main rationale for adding these exploratory results is providing a relative perspective on the model performance showed in Chapter 6, when comparing the algorithm labels with the original labels. Clearly, significant misclassification occurs among human-scorers, indicating that single-scorer data can not be treated as a perfect measure of comparison. Human-to-human agreement is more meaningful and should be regarded as the correlative gold standard of respiratory event scoring when evaluating new scoring methods. Even though the results in this section are preliminary, early speculative conclusions about the clinical implications can be made.

7.4 Clinical implications

For patients with presumed sleep disordered breathing, a PSG test helps clarify the underlying pathological mechanics and it provides an objective measure of the severity. Patients with persistent fatigue regularly show significant sleep fragmentation, often accompanied with a large AHI / RDI. While treatment of normal to mild sleep apnea (AHI < 15) focusses on behavioral therapy, moderate to severe apnea may be treated with more invasive options. The predominant occurrence of obstructive or central events both require different types of treatment methods. Treating severe obstructive apnea starts with an attempt to improve sleep hygiene, together with suggested weight loss, adjustment of sleeping positions, and optionally, oral applications. If these types of treatment appear unsuccessful, breathing assistance with CPAP may be considered. Therapy options for patients with mild central apnea are rather limited, often solely focusing on the improvement of a patient's sleep hygiene. Severe central sleep apnea can be treated with CPAP or supplemental nasal oxygen, however, many patients do not experience substantial improvement. Long-term treatment with medication remains an additional possibility, but they often come with side-effects.

Different treatment options implicate varying price-tags, which may or may not be covered by health insurance companies. Particularly in the US, this may be the decisive factor for patient to opt for cheaper treatment options, while more invasive approaches are recommended. Coverage by an insurance company is mainly based on apnea severity indicated by a PSG recording. Therefore, accurate evaluation of PSG results are vital for finding the appropriate treatment method and the according financial regulations. Inconsistent scoring of PSG recordings can have detrimental implications for a patient.

While the criteria by the AASM are based on consensus by experts in the field of sleep medicine, many of its rules are established arbitrarily. In clinical and research environments significant critique is concentrated on the fixed threshold values for the duration of flow limitations and the decrease in amplitude in the according airflow signals. Only considering flow limitations with a minimum duration of 10 seconds, instead of using an 8-second or a 12-second minimum, does not rest on clinical relevance, but is merely based on arbitrary grounds. Similarly, there is no data to suggest that a 70% or 35% would be less ore more clinically meaningful than a reduction of 90% and 30% in signal amplitude for apneas and hypopneas, respectively. In practice, clinicians often apply this minimum duration and fixed percentual amplitude decrease reluctantly when identifying apnea and hypopnea events. In addition to inconsistent scoring behavior, the difficulty to visually assess respiration signals, together with the intrinsic ambiguity on how to precisely compute flow limitations, leads to a large inter-rater variability.

Differentiation of obstructive and central events is not as pathophysiologically clear as clinical scoring may suggest. Airway collapse is common during central apnea, and high-loop gain can drive obstructive events. This biological reality of blurred boundaries leads to disagreement during

scoring, which may in part explain the significant misclassification observed in the inter-rater agreement experiment.

Thus, evaluation of a new scoring approach is difficult since the quality of manual labels is poor. Significant inter-rater variability and misclassification is observed in the gold standard of scoring, which restricts innovative data-driven approaches both in development and during performance analysis. This becomes particularly problematic when training deep learning models to perform respiratory event detection on event-level granularity. The performance analysis of the rule-based model detailed in Chapter 6 highlights the limited robustness of the AASM criteria. Assuming that the proposed algorithms adhere acceptably to the scoring criteria, the substantial number of false positives and false negatives indicate that strict compliance to the rules does not lead to optimal agreement with practitioners. Deviation from the predefined AASM thresholds would expectedly lead to the inverse effect, decreased agreement with human scorers. Instead, increasing the sensitivity of the model seems to obtain higher agreements with both included human participants.

A more sensitive scoring approach with respect to manual labelling does not necessarily result in decreased precision. Manual labelling is tedious and time-consuming, and scoring fatigue among human scorers is presumable. Therefore, the likelihood of events being missed during manual scoring is high, especially for patients with a large AHI. For these patients, often a quick diagnosis of the apnea severity can be identified, and the clinical importance to classify each single respiratory event in full night recordings deteriorates. When comparing routine manual labels to labels from an automated approach, high numbers of false positives may be the result since computer models are not possible to exhibit scoring fatigue.

An optimal balance in sensitivity and specificity performance by the model is difficult to determine. However, a strong argument can be made to prefer a highly sensitive model over its counterpart, a very specific algorithm with reduced sensitivity. Possibly, limited sensitivity leads to missing diagnoses of sleep apnea, which in turn directly restricts patients from receiving necessary treatment, whereas an overly sensitive model might suggest treatment for patients that may not (yet) require it. Expectedly, most clinicians are able to condone the latter and prefer a model with great sensitivity when screening patients for sleep apnea. In clinical setting a patient's symptoms are always considered in combination with any respiratory assessment when deciding treatment. This confines possible application of redundant therapy solely based on a potential false positive apnea diagnosis via automated PSG analysis.

The clinical advantages for automated respiratory event analysis are tremendous. Clearly, automation of PSG analysis would decrease the required analysis time and reduce costs in places where PSG analysis is already implemented. In most sleep labs, the process of manual assessment is the bottleneck that restricts the number of viable patient evaluations per day. Automation of the analysis process would allow for upscaling of the patient assessment efficacy, by multiplying the number of patient evaluations.

Moreover, automated PSG analysis computer models could be implemented in clinical centers anywhere in the world and across a variety of data acquisition options. Places without trained scorers could easily implement sleep assessment methods. Nurses and laboratory personnel would only be required to connect patients to the medical devices and oversee the measurements while all recordings can be labelled by a computer model, bypassing the most time-consuming task normally done by sleep scorers. Once automated scoring with a reduced number of input sensors is feasible, numerous additional scoring opportunities would become available including home sleep testing, testing in acute care environments, specific operational conditions such as high altitude, and consumer wearable devices. Clinically, this becomes very relevant when assessing respiratory stability and instability/events in intensive care or environmentally hostile conditions. Using limited resources, such as a respiratory effort belt, to assess respiratory abnormalities can be successfully applied in combination with other simple and small sensors necessary for monitoring patients in diverse clinical situations. Patients receiving breathing aid using CPAP would now be eligible for event detection.

8 Final thoughts

Automation approaches found in literature use different sensors and varying model architectures to detect respiratory events, which makes comparison with state-of-the-art models difficult, especially since the provided performance metrics oftentimes differ as well. Most models seem to reach a maximal global efficiency (accuracy, sensitivity and specific) within the 70-90% range, which is comparable to the human-to-human agreement observed in Chapter 7. In Chapter 5-6, I showed that both deep learning approaches and a rule-based model can compete and outperform current state-of-the-art models when globally assessing patients for apnea severity. However, differentiation between event types reduces the overall precision and leads to considerable misclassification between the event types. These findings may be explained by limited inter-rater agreement on event-level granularity, which indicates that the performance obtained with new automation approaches are confined by the provided validation data, and not only due to any inadequate model architecture or training process. The results from Chapter 7 indicate that manually scored data contains significant levels of label noise, which is related to the ambiguity and arbitrariness present in the AASM scoring guidelines that makes visual assessment of respiratory events a challenging task, as discussed in Chapter 4. In-depth assessment of inter-rater variability is unavailable in current literature, which stresses the need for a large dataset scored by multiple experts. Not only can the inter-rater variability among humans be studied, new automated approaches can be tested and verified using such a large multi-scorer dataset. Instead of using single-scorer data as the ground truth when validating models, equivalent inter-rater agreement as observed with human scorers should be pursued, and regarded as the gold-standard for validation.

The availability of multi-scorer data also creates new exciting research opportunities for the rule-based algorithms proposed in Chapter 6. Optimizing different rule-based models on individual scorers will likely lead to divergence of hyperparameter settings. Changes in hyperparameter combinations may in turn reflect the specific characteristics in signals that individual scores deem most important. In this way, scoring behavior exhibited by the human labelers can studied and variability can be quantified. Such findings will eventually help elucidating what causes inter-rater variability and may even provide support in creating more stringent rule definitions that will eventually increase scoring consistency among both automated and manual scorers.

In this work I showed that automation of respiratory event scoring is complex, yet feasible. Machine learning approaches already obtain promising results, even when reducing the input signals, but seem constrained by single-scorer data which is flawed by inherent label noise. The use of human made scoring criteria to create rule-based algorithms can obtain a similar scoring performance as deep learning approaches, but holds an important additional advantage. Automated respiratory event labelling can be achieved that closely mimics human scoring behavior, which allows for transparent hyper-parameter tuning. This makes rule-based algorithms useful tools to

study inter-rater agreement among scorers and may eventually evolve into an automated and more robust scoring method than the prevailing gold standard, manual labelling.

9 References

[1] L. Schneider, "Neurobiology and Neuroprotective Benefits of Sleep," *Contin. Lifelong Learn. Neurol.*, vol. 26, no. 4, pp. 848–870, Aug. 2020, doi: 10.1212/CON.0000000000000878.

[2] M. M. Ohayon, "Epidemiological Overview of sleep Disorders in the General Population," *Sleep Med. Res.*, vol. 2, no. 1, pp. 1–9, Apr. 2011, doi: 10.17241/smr.2011.2.1.1.

[3] A. V. 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, Aug. 2019, doi: 10.1016/S2213-2600(19)30198-5.

[4] T. L. Skaer and D. A. Sclar, "Economic implications of sleep disorders," *Pharmacoeconomics*, vol. 28, no. 11, pp. 1015–1023, 2010, doi: 10.2165/11537390-000000000-00000.

[5] J. B. Pietzsch, A. Garner, L. E. Cipriano, and J. H. Linehan, "An integrated health-economic analysis of diagnostic and therapeutic strategies in the treatment of moderate-to-severe obstructive sleep apnea," *Sleep*, vol. 34, no. 6, pp. 695–709, Jun. 2011, doi: 10.5665/SLEEP.1030.

[6] R. K. Malhotra, "Evaluating the Sleepy and Sleepless Patient," *Contin. Lifelong Learn. Neurol.*, vol. 26, no. 4, pp. 871–889, Aug. 2020, doi: 10.1212/CON.0000000000000880.

[7] H. Sun *et al.*, "Sleep staging from electrocardiography and respiration with deep learning," *Sleep*, vol. 43, no. 7, Jul. 2020, doi: 10.1093/sleep/zsz306.

[8] J. A. Dempsey, S. C. Veasey, B. J. Morgan, and C. P. O'Donnell, "Pathophysiology of sleep apnea," *Physiological Reviews*, vol. 90, no. 1. American Physiological Society, pp. 47–112, 2010, doi: 10.1152/physrev.00043.2008.

[9] S. Javaheri *et al.*, "Sleep Apnea: Types, Mechanisms, and Clinical Cardiovascular Consequences," *Journal of the American College of Cardiology*, vol. 69, no. 7. Elsevier USA, pp. 841–858, Feb. 21, 2017, doi: 10.1016/j.jacc.2016.11.069.

[10] M. S. Avidan *et al.*, "Obstructive sleep apnea as an independent predictor of postoperative delirium and pain: Protocol for an observational study of a surgical cohort [version 2; referees: 2 approved]," *F1000Research*, vol. 7, 2018, doi: 10.12688/f1000research.14061.2.

[11] S. L. Revels, B. H. Cameron, and R. B. Cameron, "Obstructive sleep apnea and perioperative delirium among thoracic surgery intensive care unit patients: Perspective on the STOP-BANG questionnaire and postoperative outcomes," *J. Thorac. Dis.*, vol. 11, no. Suppl 9, pp. S1292–S1295, 2019, doi: 10.21037/jtd.2019.04.63.

[12] M. T. Naughton, "Cheyne-stokes respiration," *Sleep Medicine Clinics*, vol. 9, no. 1. W.B. Saunders, pp. 13–25, Nov. 20, 2014, doi: 10.1016/j.jsmc.2013.11.002.

[13] S. Javaheri and J. A. Dempsey, "Central sleep apnea," *Compr. Physiol.*, vol. 3, no. 1, pp. 141–163, 2013, doi: 10.1002/cphy.c110057.

[14] M. Arzt *et al.*, "Suppression of central sleep apnea by continuous positive airway pressure and transplant-free survival in heart failure: A post hoc analysis of the Canadian Continuous Positive Airway Pressure for Patients with Central Sleep Apnea and Heart Failure Trial (CANPAP)," *Circulation*, vol. 115, no. 25, pp. 3173–3180, Jun. 2007, doi: 10.1161/CIRCULATIONAHA.106.683482.

[15] S. Javaheri, "Effects of continuous positive airway pressure on sleep apnea and ventricular irritability in patients with heart failure," *Circulation*, vol. 101, no. 4, pp. 392–397, Feb. 2000, doi:

10.1161/01.CIR.101.4.392.

[16] Meir H. Kryger; Thomas Roth; William C. Dement, *Principles and Practice of Sleep Medicine*. Elsevier, 2017.

[17] L. Vargas-Ramirez, M. Gonzalez-Garcia, C. Franco-Reyes, and M. A. Bazurto-Zapata, "Severe sleep apnea, Cheyne-Stokes respiration and desaturation in patients with decompensated heart failure at high altitude," *Sleep Sci.*, vol. 11, no. 3, pp. 146–151, 2018, doi: 10.5935/1984-0063.20180028.

[18] Y. Kim, S. Kim, D. R. Ryu, S. Y. Lee, and K. Bin Im, "Factors associated with Cheyne-stokes respiration in acute ischemic stroke," *J. Clin. Neurol.*, vol. 14, no. 4, pp. 542–548, Oct. 2018, doi: 10.3988/jcn.2018.14.4.542.

[19] A. Ogna *et al.*, "Prevalence and clinical significance of respiratory effort-related arousals in the general population," *J. Clin. Sleep Med.*, vol. 14, no. 8, pp. 1339–1345, Aug. 2018, doi: 10.5664/jcsm.7268.

[20] J. L. Pépin, M. Guillot, R. Tamisier, and P. Lévy, "The Upper Airway Resistance Syndrome," *Respiration*, vol. 83, no. 6, pp. 559–566, Jun. 2012, doi: 10.1159/000335839.

[21] C. Cracowski, J. L. Pépin, B. Wuyam, and P. Lévy, "Characterization of obstructive nonapneic respiratory events in moderate sleep apnea syndrome," *Am. J. Respir. Crit. Care Med.*, vol. 164, no. 6, pp. 944–948, Sep. 2001, doi: 10.1164/ajrccm.164.6.2002116.

[22] A. van den Oord *et al.*, "WaveNet: A Generative Model for Raw Audio," Sep. 2016, Accessed: Aug. 10, 2020. [Online]. Available: http://arxiv.org/abs/1609.03499.

[23] M. Zabihi, A. B. Rad, S. Kiranyaz, S. Särkkä, and M. Gabbouj, "1D Convolutional Neural Network Models for Sleep Arousal Detection," Mar. 2019, Accessed: Aug. 10, 2020. [Online]. Available: http://arxiv.org/abs/1903.01552.

[24] S. S. Mostafa, F. Mendonça, A. G. Ravelo-García, and F. Morgado-Dias, "A systematic review of detecting sleep apnea using deep learning," *Sensors (Switzerland)*, vol. 19, no. 22. MDPI AG, Nov. 02, 2019, doi: 10.3390/s19224934.

[25] 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, Mar. 2018, doi: 10.1088/1361-6579/aaafb8.

[26] 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, doi: 10.1109/TITB.2010.2087386.

[27] T. Van Steenkiste, W. Groenendaal, Di. Deschrijver, and T. Dhaene, "Automated Sleep Apnea Detection in Raw Respiratory Signals Using Long Short-Term Memory Neural Networks," *IEEE J. Biomed. Heal. Informatics*, vol. 23, no. 6, pp. 2354–2364, Nov. 2019, doi: 10.1109/JBHI.2018.2886064.

[28] R. Haidar, S. McCloskey, I. Koprinska, and B. Jeffries, "Convolutional Neural Networks on Multiple Respiratory Channels to Detect Hypopnea and Obstructive Apnea Events," in *Proceedings of the International Joint Conference on Neural Networks*, Oct. 2018, vol. 2018-July, doi: 10.1109/IJCNN.2018.8489248.

[29] S. Biswal, H. Sun, B. Goparaju, M. Brandon Westover, J. Sun, and M. T. Bianchi, "Expert-level sleep scoring with deep neural networks," *J. Am. Med. Informatics Assoc.*, vol. 25, no. 12, pp.

1643–1650, Dec. 2018, doi: 10.1093/jamia/ocy131.

[30] N. Banluesombatkul, T. Rakthanmanon, and T. Wilaiprasitporn, “Single Channel ECG for Obstructive Sleep Apnea Severity Detection using a Deep Learning Approach,” *IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON*, vol. 2018-Octob, pp. 2011–2016, Aug. 2018, doi: 10.1109/TENCON.2018.8650429.

[31] P. Lakhan, A. Ditthapron, N. Banluesombatkul, and T. Wilaiprasitporn, “Deep Neural Networks with Weighted Averaged Overnight Airflow Features for Sleep Apnea-Hypopnea Severity Classification,” *IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON*, vol. 2018-Octob, pp. 441–445, Aug. 2018, doi: 10.1109/TENCON.2018.8650491.

[32] F. Espinoza-Cuadros, R. Fernández-Pozo, D. T. Toledano, J. D. Alcázar-Ramírez, E. López-Gonzalo, and L. A. Hernández-Gómez, “Speech Signal and Facial Image Processing for Obstructive Sleep Apnea Assessment,” *Comput. Math. Methods Med.*, vol. 2015, 2015, doi: 10.1155/2015/489761.

[33] G. Gutiérrez-Tobal, D. Álvarez, J. Gomez-Pilar, F. del Campo, and R. Hornero, “Assessment of Time and Frequency Domain Entropies to Detect Sleep Apnoea in Heart Rate Variability Recordings from Men and Women,” *Entropy*, vol. 17, no. 1, pp. 123–141, Jan. 2015, doi: 10.3390/e17010123.

[34] T. Rosenwein, E. Dafna, A. Tarasiuk, and Y. Zigel, “Breath-by-breath detection of apneic events for OSA severity estimation using non-contact audio recordings,” in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, Nov. 2015, vol. 2015-Novem, pp. 7688–7691, doi: 10.1109/EMBC.2015.7320173.

[35] T. Kim, J. W. Kim, and K. Lee, “Detection of sleep disordered breathing severity using acoustic biomarker and machine learning techniques,” *Biomed. Eng. Online*, vol. 17, no. 1, Feb. 2018, doi: 10.1186/s12938-018-0448-x.

[36] S. McCloskey, R. Haidar, I. Koprinska, and B. Jeffries, “Detecting hypopnea and obstructive apnea events using convolutional neural networks on wavelet spectrograms of nasal airflow,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Jun. 2018, vol. 10937 LNAI, pp. 361–372, doi: 10.1007/978-3-319-93034-3_29.

[37] S. H. Choi *et al.*, “Real-time apnea-hypopnea event detection during sleep by convolutional neural networks,” *Comput. Biol. Med.*, vol. 100, pp. 123–131, Sep. 2018, doi: 10.1016/j.compbiomed.2018.06.028.

[38] R. Haidar, I. Koprinska, and B. Jeffries, “Sleep apnea event detection from nasal airflow using convolutional neural networks,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017, vol. 10638 LNCS, pp. 819–827, doi: 10.1007/978-3-319-70139-4_83.

[39] D. Álvarez, R. Hornero, D. Abásolo, F. Del Campo, and C. Zamarrón, “Nonlinear characteristics of blood oxygen saturation from nocturnal oximetry for obstructive sleep apnoea detection,” *Physiol. Meas.*, vol. 27, no. 4, pp. 399–412, Apr. 2006, doi: 10.1088/0967-3334/27/4/006.

[40] H. ElMoaqet, J. Kim, D. Tilbury, S. K. Ramachandran, M. Ryalat, and C.-H. Chu, “Gaussian Mixture Models for Detecting Sleep Apnea Events Using Single Oronasal Airflow Record,” *Appl. Sci.*, vol. 10, no. 21, p. 7889, Nov. 2020, doi: 10.3390/app10217889.

[41] N. A. Collop, “Scoring variability between polysomnography technologists in different sleep

laboratories," *Sleep Med.*, vol. 3, no. 1, pp. 43–47, 2002, doi: 10.1016/S1389-9457(01)00115-0.

[42] S. T. Kuna *et al.*, "Agreement in computer-assisted manual scoring of polysomnograms across sleep centers," *Sleep*, vol. 36, no. 4, pp. 583–589, Apr. 2013, doi: 10.5665/sleep.2550.

[43] H. Elmoaqet, M. Eid, M. Glos, M. Ryalat, and T. Penzel, "Deep recurrent neural networks for automatic detection of sleep apnea from single channel respiration signals," *Sensors (Switzerland)*, vol. 20, no. 18, pp. 1–19, Sep. 2020, doi: 10.3390/s20185037.

[44] J. A. Reichert, D. A. Bloch, E. Cundiff, and B. Votteri, "Comparison of the NovaSom QCG™, a new sleep apnea home-diagnostic system, and polysomnography," *Sleep Med.*, vol. 4, no. 3, pp. 213–218, 2003, doi: 10.1016/S1389-9457(02)00234-4.

[45] N. Scalzitti, S. Hansen, S. Maturo, J. Lospinoso, and P. O'Connor, "Comparison of home sleep apnea testing versus laboratory polysomnography for the diagnosis of obstructive sleep apnea in children," *Int. J. Pediatr. Otorhinolaryngol.*, vol. 100, pp. 44–51, Sep. 2017, doi: 10.1016/j.ijporl.2017.06.013.

[46] S. Su, F. M. Baroody, M. Kohrman, and D. Suskind, "A comparison of polysomnography and a portable home sleep study in the diagnosis of obstructive sleep apnea syndrome," *Otolaryngol. - Head Neck Surg.*, vol. 131, no. 6, pp. 844–850, Dec. 2004, doi: 10.1016/j.otohns.2004.07.014.

[47] M. Howe-Patterson, B. Pourbabaee, and F. Benard, "Automated Detection of Sleep Arousals from Polysomnography Data Using a Dense Convolutional Neural Network," in *Computing in Cardiology*, Sep. 2018, vol. 2018-September, doi: 10.22489/CinC.2018.232.

[48] M. Friedman, *Sleep Apnea and Snoring*. Elsevier, 2009.

Automated Scoring of Respiratory Events in Sleep with a Single Effort Belt and Deep Neural Networks

Thijs E Nassi, Wolfgang Ganglberger, Haoqi Sun, Abigail A Bucklin, Siddharth Biswal, Michel J A M van Putten, Robert J Thomas, M Brandon Westover

***Abstract*— The gold standard to assess respiration during sleep is polysomnography; a technique that is burdensome, expensive (both in analysis time and measurement costs), and difficult to repeat. Automation of respiratory analysis can improve test efficiency and enable accessible implementation opportunities worldwide. Using 9,656 polysomnography recordings from the Massachusetts General Hospital (MGH), we trained a neural network (WaveNet) based on a single respiratory effort belt to detect obstructive apnea, central apnea, hypopnea and respiratory-effort related arousals. Performance evaluation included event-based and recording-based metrics – using an apnea-hypopnea index analysis. The model was further evaluated on a public dataset, the Sleep-Heart-Health-Study-1, containing 8,455 polysomnographic recordings. For binary apnea event detection in the MGH dataset, the neural network obtained an accuracy of 96%, an apnea-hypopnea index r^2 of 0.90 and area under the curve for the receiver operating characteristics curve and precision-recall curve of 0.93 and 0.71, respectively. For the multiclass task, we obtained varying performances: 84% of all labeled central apneas were correctly classified, whereas this metric was 51% for obstructive apneas, 40% for respiratory effort related arousals and 23% for hypopneas. The majority of false predictions were misclassifications as another type of respiratory event. Our fully automated method can detect respiratory events and assess the apnea-hypopnea index with sufficient accuracy for clinical utilization. Differentiation of event types is more difficult and may reflect in part the complexity of human respiratory output and some degree of arbitrariness in the clinical thresholds and criteria used during manual annotation.**

***Index Terms*— Sleep apnea, Respiratory event detection, Respiratory effort, Deep learning, Apnea Hypopnea Index, Polysomnography**

I. Introduction

Sleep disorders such as sleep apnea and insomnia affect millions of people worldwide [1]. Clinical effects include difficulty in initiating and maintaining sleep, impaired alertness, and hypertension. Excessive daytime sleepiness and fatigue, two common symptoms associated with sleep disorders, have a large impact on population health [2], [3]. Accurate and timely diagnosis of a patient's sleep disorder is therefore essential. Patients with apnea, especially obstructive sleep apnea, are at increased risk for traffic accidents, postoperative complications, and delirium [4], [5]. Untreated sleep apnea is associated with arrhythmias, heart failure and stroke. Studies that measure the apnea-hypopnea index (AHI) show that an estimated 49.7% of male and 23.4% of female adults have moderate-to-severe sleep-disordered breathing, though a lower percentage are clinically symptomatic [5].

The gold standard to measure sleep objectively is laboratory-based polysomnography (PSG). PSG is conventionally scored based on the American Academy of Sleep Medicine (AASM) guidelines. Scoring PSG recordings is a time-consuming task performed by specialists in dedicated sleep centers, making this an expensive process both in time and costs. Automation of PSG analysis would decrease the required analysis time and reduce costs. Moreover, automated PSG analysis computer models could be implemented in clinical centers anywhere in the world and across a variety of data acquisition options, including home sleep testing, testing in acute care environments, specific operational conditions such as high altitude, and consumer wearable devices.

Medical data is complex and involves a large number of variables and context that are difficult to encompass by programs based on a fixed set of rules. Deep learning models such as convolutional neural networks (CNN) and recurrent neural networks (RNN) have been applied in many domains to solve complex pattern recognition tasks [6]. Deep learning algorithms rely on patterns and inference rather than explicit instructions and can learn intricate relationships between features and labels from data. Implementing neural networks has become relevant in analyzing the heterogeneous kinds of data generated in modern clinical care [7]. Various types of deep learning algorithms have been found to be suitable for analyzing specific types of data. For instance, CNNs have been successful in classifying objects in images. Typical CNN architectures, however, are not ideal when analyzing temporal data. Temporal data is typically better exploited by RNNs. However, the recently introduced CNN, WaveNet architecture has been found to perform better than RNNs on several tasks [8]. WaveNet's architecture resembles a typical CNN, yet the application of dilated causal convolutions creates an effectively larger receptive field. This renders WaveNet capable of detecting both spatial patterns and long-range temporal patterns. WaveNet was originally designed to synthesize speech; however, its application has been found suitable for analyzing other types of signals. In 2018 a challenge organized by the PhysioNet Computing in Cardiology aimed to detect sleep arousals from a variety of physiological signals, including signals derived from respiration. The winning model was a modified WaveNet architecture, suggesting that this CNN architecture can indeed perform successfully in other domains such as the automation of PSG-related tasks [9].

In the last two years a significant number of papers have been published on the detection of sleep apnea, as described by recent review papers [24], [25]. Finding a patient-friendly and accurate sensor or signal, especially in combination with a suitable analysis model, is clearly an ongoing area of high relevance. An overview of

M.B.W. Was supported by the Glenn Foundation for Medical Research and American Federation for Aging Research (Breakthroughs in Gerontology Grant); American Academy of Sleep Medicine (AASM Foundation Strategic Research Award); Football Players Health Study (FPHS) at Harvard University; Department of Defense through a subcontract from Moberg ICU Solutions, Inc; by the NIH (1R01NS102190, 1R01NS102574, 1R01NS107291, 1RF1AG064312). (Corresponding author: M. Brandon Westover.)

T.E. Nassi and M.J.A.M Van Putten are with University of Twente, 7522NB Enschede, the Netherlands (e-mail: t.nassi@student.utwente.nl; m.j.a.m.vanputten@utwente.nl).

W. Ganglberger, H. Sun, A.A. Bucklin and M.B. Westover are with Massachusetts General Hospital, Boston, MA, 02114 USA (e-mail: wganglberger@mgh.harvard.edu; abucklin@partners.org; mwestover@mgh.harvard.edu).

S. Biswal is with School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA (e-mail: siddnitr1@gmail.com).

R.J. Thomas is with Deaconess Medical Center, Boston, MA, 02215, USA (e-mail: rthomas1@bidmc.harvard.edu).

TABLE I
OVERVIEW OF OTHER STUDIES PERFORMING AUTOMATED RESPIRATORY EVENT DETECTION USING 96 PATIENTS OR MORE

Study	Dataset size	Signal type	Analysis model	Classifier type	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-score (%)	AUC_{ROC} (%)	AUC_{PR} (%)
[10]	10.000	Airflow, respiration chest, abdomen, oxygen saturation	RCNN	G	88.2	-	-	-	-	-	-
[11]	2100	Respiration abdomen	LSTM	A/N	77.2	62.3	80.3	39.9	-	77.5	45.3
[12]	1507	Nasal airflow, abdominal, thoracic plethysmography	CNN1D-3ch		83.5	83.4	-	83.4	83.4	-	-
[13]	1507	Nasal airflow	CNN2D	A/H/N	79.8	79.9	-	79.8	79.7	-	-
[14]	545	Electrocardiography	CNN1D-LSTM MHLNN	G	79.5	77.6	80.1	-	79.1	-	-
[15]	520	Airflow	MHLNN	G	87.2	88.3	87.8	-	-	-	-
[16]	285	Voice and facial features	GMM	G	72	73	65	-	-	-	-
[17]	188	Airflow, respiratory rate variability	LR	G	72	80	59	-	-	-	-
[18]	187	pulse oximetry	CTM	G	87	90	83	-	-	-	-
[19]	186	Breathing sounds	Binary-RF	G	86	-	-	-	-	-	-
[20]	179	Nasal pressure	CNN1D	A/H/N	96.6	81.1	98.5	87	-	-	-
[21]	120	Breathing sounds	MHLNN	G	75	-	-	-	-	-	-
[22]	100	Nasal airflow	CNN1D	OA/N	74.7	74.7	-	74.5	-	-	-
[23]	96	Nasal airflow	GMM	OA/CA/H/N	83.4	88.5	82.5	46.6	42.7	86.7	-

Analysis models: RCNN = recurrent and convolutional neural networks, LSTM = long short-term memory, CNN = convolution neural network, MHLNN = multiple hidden layers neural network, GMM = gaussian mixture model, LR = logistic regression, CTM = central tendency measure, RF = random forest. Classifier types: A = apnea, H = hypopnea, N = normal, O = obstructive, C = central G = global.

other sleep apnea studies that use large datasets (at least 96 patients) can be found in Table I.

Sleep apnea detection methods typically use various breathing measurements and oximetry [25]. Alternative methods using signals derived from electrocardiography (ECG) have shown some promise for predicting AHI as well, although such data has an indirect relationship to the respiratory system and therefore to sleep apnea [24], [26]. This more indirect method of analyzing respiration requires additional processing and can be affected by other illnesses including heart failure and cardiac arrhythmias, rather than sleep apnea [11]. Classification of respiratory events typically requires both airflow and respiratory effort signals. Using multiple physiological signals to detect sleep apnea can provide good performance [10], [22]. However, this leads to similar problems as the current gold standard; using many different sensor signals is considered uncomfortable, expensive, and time-consuming. Recent studies show that automated apnea scoring with limited sensors use (i.e. airflow or respiratory effort) can still yield acceptable performance [11], [23], [27]. Using airflow or respiratory effort for apnea detection bear different advantages and disadvantages. Airflow measures are expected to yield slightly better performance but need access to the nose/mouth, which may be difficult in specific environments. In situations where the airflow signal may not be readily acquired, an effort-belt based classification could overcome this limitation. Examples include in intensive care units, home tracking in heart failure or chronic obstructive pulmonary disease, those using nasal oxygen, and war fighter conditions. The effort belt is highly convenient, and this input signal can be acquired by a range of contact and contactless technologies in nearly every possible environment.

The ability to identify and discriminate between the specific respiratory events that are typically scored in PSG while using fewer signals is unknown to the current clinical setting. In this research we aimed to create a fully automated method that can detect respiratory events, discriminate between the different types of respiratory events, and assess the AHI with sufficient efficiency for clinical implementation using only a single respiratory effort belt.

II. METHODS

A. Dataset

The dataset used to train our model was from The Massachusetts General Hospital (MGH) sleep laboratory (2008-2018), summarized in Table II. The MGH Institutional Review Board approved the retrospective analysis of the clinically acquired PSG data. In total 9656 PSG recordings were successfully exported. We applied a 5-fold cross-validation for which we split the dataset into training, validation, and test subsets using respective ratios of 80%, 10%, 10% of the total number of recordings. Multiple records from the same patients were constrained to the same fold. Patients with and without breathing assistance by continuous positive airway pressure (CPAP) were included.

We included a secondary dataset for external validation of our model. This dataset was collected by the Sleep Heart Health Study (SHHS) and included 8455 PSG recordings. For this research we only used the signal measured at the abdomen using a respiratory effort belt (inductance plethysmography). This signal, in comparison to the available respiratory signals measured on the thorax, is expected to provide the best predictive performance [28].

The MGH sleep center is an AASM accredited sleep center, with stringent ongoing requirements for documenting and maintaining high inter-scorer reliability. The center maintains an inter-rater reliability of over 85%. Respiratory event detections included obstructive apneas, central apneas, mixed apneas, hypopneas, and respiratory effort-related arousals (RERAs). Because of the relatively low number of mixed apnea events in our dataset, 1.7% of all events, all mixed apnea events were labeled as obstructive apnea, since the characteristics are expected to look most similar. We define respiratory events as a term that encompasses any type of apnea, hypopnea and RERA. The definition apnea reflects any type of apnea and hypopnea.

Recordings obtained from the SHHS database were annotated according to SHHS guidelines. A key difference between the two datasets is the primary respiratory scoring signal in the original source – nasal pressure (MGH) and thermistor (SHHS). This difference and implications will be discussed further below. Besides the different flow sensors the MGH and SHHS dataset only include labels that are scored using the same criteria as defined by the AASM (4% rule for hypopneas), and individual recordings were annotated by a single scorer for both datasets. For the MGH data there was a total of 7 scorers whereas for the SHHS data the number of experts is not reported. We chose the SHHS dataset as it was the largest study which used a uniform methodology for acquisition and scoring.

B. Preprocessing and data preparation

All recordings that were incomplete or did not include any sleep were removed. For the SHHS dataset, we only used recordings that contain mostly good quality abdominal effort signals, as defined by the SHHS [29]. Specifically, for the visit 1 and visit 2 subsets, only recordings with at least 4 hours of artifact-free signal or 75% of artifact-free signal, respectively, were included. To extract the relevant respiratory information and remove present noise, minimal preprocessing techniques were applied. The abdominal respiration measurement from both the MGH data and the SHHS data consisted of a single channel with a sampling frequency of 125 Hz, 200 Hz or 250 Hz. A notch filter of 60 Hz was applied to reduce line noise. A low-pass filter of 10 Hz was applied to remove higher frequencies not of interest, and consequently all recordings were resampled to 10 Hz. Z-score normalization was performed using the mean and standard deviation of the 1^{st} to 99^{th} percentile clipped signal to optimize the training process of the neural network.

The training data was segmented into 7-minute segments with a stride of 30 seconds to reduce the large training dataset size. Each segment was assigned one ground truth class label – the sleep expert's label located in the center of the segment. We segmented the test data in the same way, except that we used a stride of 1 second, which allowed for a respiratory event prediction for each second.

C. Model and prediction tasks

In this research we utilized a WaveNet model (see model architecture in Section II-D) to automatically detect apneas, hypopneas, and RERAs from a single effort belt signal, without use of additional sensors that are conventional in PSG measurements (e.g. thermistor, nasal pressure, oxygen saturation, electroencephalography or electrocardiography), and without using human-engineered features. As described above, the signal was split into 7-minute segments and, in this way, the model was trained to predict only the center index of a 7-minute segment, while having 3.5 minutes of context information before and after the center index. We designed the following two prediction tasks:

- Binary classification to discriminate non-apnea events from apnea-hypopnea events (regular breathing and respiratory events). Based on the predicted respiratory events, we computed the predicted AHI as the number of predicted respiratory events per hour of sleep.
- Multiclass classification to discriminate the respiratory event classes: no-event, obstructive apnea or mixed apnea, central apnea, RERA, hypopnea. From the sum of the detected respiratory events, we determined the AHI and respiratory disturbance index (RDI).

In both tasks we used the originally scored multiclassification labels. For our binary classification task we converted all types of apnea and hypopneas into one grouped class, apnea. In both experiments our model provided a probability for all included classes. The highest probability among the possible classes constitutes the output of our model. In Fig. 1 the complete workflow scheme is shown.

TABLE II
DATASET DISTRIBUTION (N=9656)

Category	Bin	Percentage of all patients
Sex	male	58.9%
	female	40.7%
	unknown	0.4%
Age	< 60	65.0%
	60 - 80	34.4%
	> 80	2.6%
BMI	underweight (< 18.5)	0.9%
	normal weight (18.5 - 25)	13.7%
	overweight (25 - 30)	26.7%
	obese (> 30)	58.7%
AHI	normal (< 5)	39.9%
	mild (5 - 15)	27.7%
	moderate (15 - 30)	20.4%
	severe ($>= 30$)	12.0%
Recording type	diagnostic	48.1%
	split night	24.3%
	all night CPAP	24.0%
	unknown	3.6%
Events (N=675,667)	obstructive apnea	18.6%
	central apnea	14.2%
	mixed apnea	1.7%
	hypopnea	36.8%
	RERA	28.7%

D. Model architecture

WaveNet is a fully convolutional neural network [8], [30]. In Appendix Fig. 7 we show the schematics of a residual block of the WaveNet model. The architecture makes use of an exponentially increasing dilation factor resulting in exponential growth of the receptive field with each layer. This causes the receptive field to double in length for each hidden layer. In previous work we showed that 4.5-minute segments are ideal for sleep staging from respiratory effort data [31]. The exponential growth of receptive field gives us a limited number of options without making major changes to the fundamental architecture of the WaveNet model (10 layers is equivalent to 1.7 minutes, 11 layers is equivalent to 3.4 minutes, and 12 layers is equivalent to 6.8 minutes). To ensure enough context for our respiratory event scoring task we opted for 12 hidden layers, resulting in 4096 samples in our 10 Hz signal, equivalent to approximately 7 minutes of context. Instead of using WaveNet as a generative model that uses the last output as its subsequent input (recurrent generation), we trained WaveNet in a supervised manner where the input is a respiratory effort signal and the prediction target is either binary or quinary, for experiment 1 and 2 respectively. The original WaveNet model makes use of causal convolutions where the output is a function of previous time steps only, no future time steps. For this research we modified the WaveNet architecture by using non-causal convolutions and shifting the output node, which results in that the output is now a function of both previous and future time steps. Non-causal connection, i.e. past and future context, matches better to a human sleep scorer conceptually, as they have access to the full night recording. The number of filters for each of the convolutions was set to 32. A kernel size of 2 was used. The categorical cross entropy loss function was applied during training,

$$\text{loss} = \sum_{i=1}^{N} -y_i' log(y_i), \tag{1}$$

where y represents the predicted probability distribution, $y\prime$ represents the true distribution, and N represents the number of classes. To address overfitting and to improve generalization of the network, besides using a dropout rate of 0.2 we have implemented an early stopping procedure, where we stop training if the performance on a validation set does not increase for 10 consecutive training rounds. We used a batch size of 150 segments and a learning rate of 0.001, which was reduced to 10% when three consecutive training rounds showed no improvement. ADAM optimization was used for training the classifier.

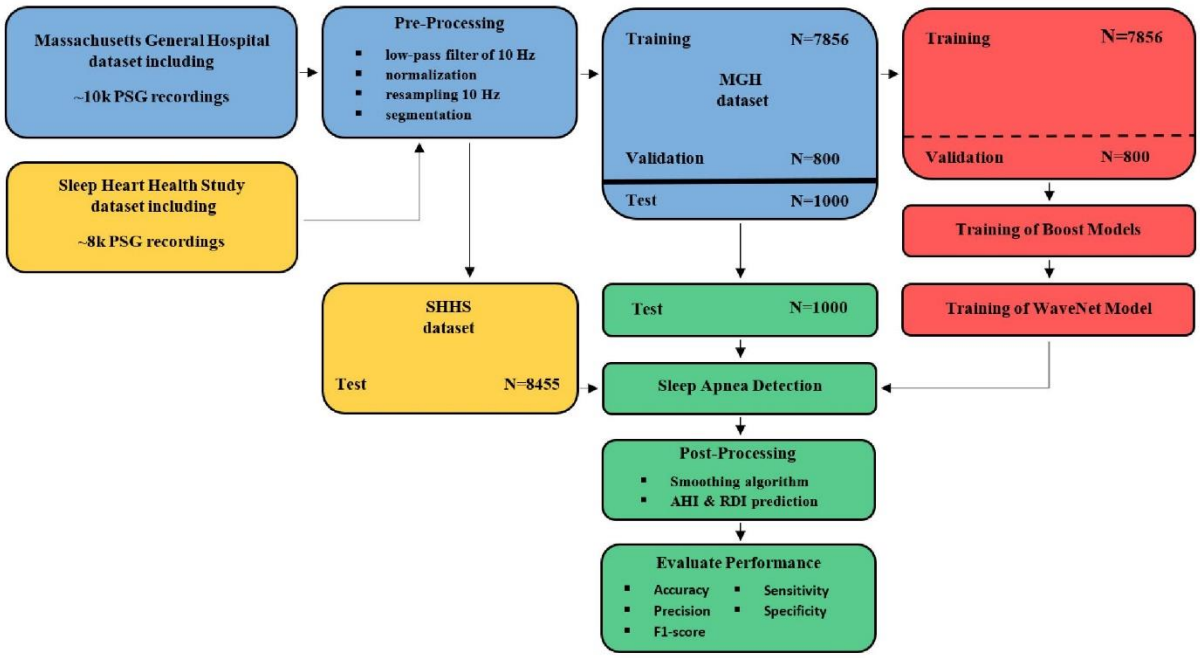


Fig. 1. Data flow scheme for model development and testing. The model was trained and validated on the dataset from the Massachusetts General Hospital (MGH) whereas the dataset from the Sleep Heart Health Study (SHHS) was used for external validation. Both the Apnea Hypopnea Index (AHI) and the respiratory disturbance index (RDI) were computed during post-processing.

E. Boosting for imbalanced data

Classification with imbalanced data is challenging in many real-world deep learning applications [32], [33]. For the PSG recordings in our research, the number of segments containing only regular breathing is typically much larger than the segments containing respiratory events, even for patients classified with severe apnea. For this problem we designed a boosted model approach by applying a binary WaveNet classifier, or boost-model, over multiple iterations. To remove a large proportion of segments with regular breathing without removing many segments including apnea events, only segments with an extremely high probability of regular breathing were removed by the boost-model. In our approach we selected a probability threshold to make our boost-model extremely sensitive for apneas, based on the receiver operating characteristic (ROC) curve. In the first iteration we used a true positive rate of 0.995 and decreased this value by 0.010 for each subsequent iteration. The boosted model iterations stopped when the desired balance in classes was obtained. This balance was defined by 3:1 ratio of regular breathing with respect to the sum of events in experiment 1, and a 3:1 ratio of regular breathing with respect to the most commonly occurring event type in experiment 2. In Fig. 2 the boosted model flow scheme is shown.

In each iteration, the boost model received non-rejected samples from the previous iteration. Using this approach, the boost-model was trained to discriminate regular breathing from other respiratory events, while being exposed to a decreasing and increasingly more challenging dataset. In this way, in every iteration our boost-model should learn new nuanced characteristics that define a normal breathing rhythm. Using this boosted approach, we vastly reduced the number of segments containing regular breathing and improved effective classification by our main WaveNet model. Moreover, the boost-model was expected to remove segments with regular breathing that are relatively simple to distinguish from apnea while leaving the more complex segments for our main model. All boost-models were trained on the training dataset using the same hyperparameters as our main binary model, and were applied on the test dataset.

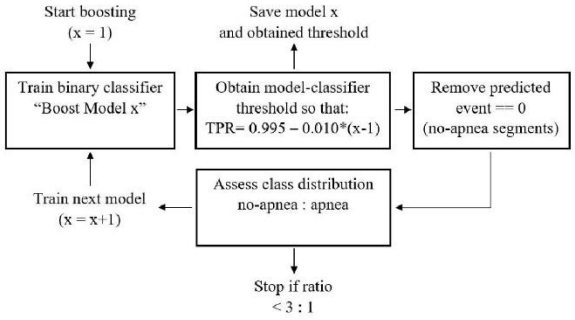


Fig. 2. Flow scheme for the boosted model process that was performed to create a more balanced dataset to train our WaveNet model.

F. Post-processing

After applying our model, we obtained an apnea prediction for each second. The prediction resolution of 1 Hz allowed high fluctuations of predicted events and allowed detection of very brief events. Both situations were considered not physiologically plausible, therefore, we designed a smoothing algorithm. This smoothing algorithm removed short events and rapidly changing events. The algorithm was based on a moving window of 10 seconds. The following rules were applied for each window.

- When a minimum of 3 out of 10 seconds was classified as no-event, the complete window was set to no-event.

- If a window was classified as an event and multiple types of events were present, the type of event that occurred most became the prediction for the complete window.

The selection of 3 seconds was based on manual, visual analysis on classifier outputs on a small subset of the patients in the training set (less than 20 patients). We do not believe the performance is sensitive to the choice of this parameter (i.e. between 2 and 4 out of the 10 seconds).

Finally, consecutive events with a combined length of two windows or greater were converted into a single event prediction. The type of event that held the largest proportion among the combined events indicated this new prediction. Applying the smoothing algorithm resulted in predicted events with a minimum length of 10 seconds, similar to what is suggested by AASM guidelines.

To obtain a more clinically interpretable performance granularity, we specified the following criterion to judge when a detected event overlapped sufficiently with an event annotated by sleep experts to count as correct:

- A predicted event is considered a correct prediction when more than 50% of its duration overlaps with an expert label.

G. Model evaluation

Confusion matrices were computed to assess per-event performance of our model. The true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) were computed for each of the 2 or 5 classes, respectively, for experiment 1 and 2. To obtain a TN value of similar quantization as positive events, i.e. per event granularity, we used a data-driven way of obtaining the duration of negative apnea events. This was accomplished by accumulating the time where no event was found by neither the experts nor by our model and dividing this duration by the median length of all respiratory events in the MGH test data, i.e. 18 seconds. Next, the TP, TN, FP and FN values were used to determine the following event-per-event performance metrics of our model.

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

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

$$\text{F1 score} = \frac{2\times\text{Sensitivity}\times\text{Precision}}{\text{Sensitivity}+\text{Precision}}$$

Additionally, Cohen's kappa values were determined using the formula in Cohen's original work [34]. For all above mentioned metrics, we obtained the 95% confidence interval by bootstrapping over patients (sampling with replacement by blocks of patients) 10.000 times. The confidence interval was computed as the 2.5% (lower bound) and the 97.5% percentile (upper bound). For the binary tasks the ROC curve and precision-recall (PR) curve and their corresponding areas (AUC_{ROC} AUC_{PR}) were computed.

In addition to event-per-event evaluation, we evaluated global scoring performance. Global assessment of sleep apnea severity is typically used for clinical diagnosis [24]. For the first experiment we determined the AHI value per patient, whereas in the second experiment we determined the AHI and RDI value per patient using the following computations.

$$\text{AHI} = \frac{\text{OA}+\text{CA}+\text{HY}}{\text{hours of sleep}}$$

$$\text{RDI} = \frac{\text{OA}+\text{CA}+\text{HY}+\text{RERAs}}{\text{hours of sleep}}$$

where OA is obstructive apneas, CA is central apneas, and HY is hypopneas. We used the already scored sleep stages from the original annotations to differentiate between sleep and wake time for the patients. We decide to not use "time in bed" as the denominator, as in previous work we have shown it is possible to reasonably stage sleep with one effort belt signal as input [31]. To keep the focus of this paper on apnea detection with one effort belt, we believe using the expert sleep labels helps to best answer our research question.

With the AHI score all patients were categorized as normal or mild, moderate or severe sleep apnea. Categorization was according to conventional criteria as defined by AASM guidelines.

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

We obtained the classification accuracy of our model by creating a confusion matrix for the four AHI scores. The classification accuracy displays the ability of the model to assign a patient to any of the four AHI categories. To gain insight into accuracy of the AHI prediction disregarding the discrete borders used in categorization, histograms were computed to show the difference between the AHI value scored by the experts and the AHI value predicted by our model. For both experiments, scatter plots visualizing the correlation between the expert-scored AHI and the algorithm-predicted AHI were computed. Additionally, for experiment 2, we computed scatter plots for the RDI and each type of respiratory event per hour of sleep. A robust linear regression model with bi-squared cost function was fitted to the data to compute the correlation between the scored AHI by the experts and predicted AHI by our model [35]. This model was selected to mitigate the effect of outliers. Also, Cohen's kappa values were determined for AHI prediction.

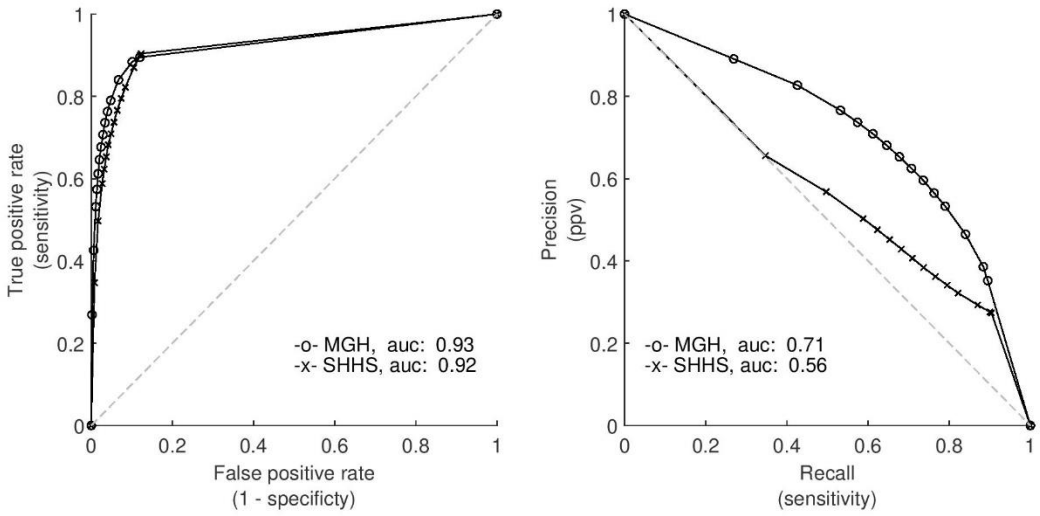
III. Results

For the MGH and SHHS testing dataset, 16 and 1128 recordings, respectively, were removed due to insufficient sleep or erroneous data. The boosted model approach resulted in 5 consecutive model iterations before reaching the desired class balance in both experiment 1 and 2. The median length among all respiratory events was 18 seconds. This length was used to determine the number of true negative events. In Appendix Fig. 8 an example recording can be found. As the large flat parts show no continuous false positive predictions we are convinced that our model learned not to classify such regions as respiratory events.

A. Per-event performance

An overview of the per-event performance metrics for the binary task are given in Table III. Both absolute and normalized confusion matrices for MGH and SHHS dataset are shown in Appendix Table VI. An AUC_{ROC} value of 0.93 and AUC_{PR} of 0.71 was found for the MGH dataset. The AUC_{ROC} and AUC_{PR} for the SHHS dataset were 0.92 and 0.56 (see Fig. 3 for the ROC and PR curves). Appendix Fig. 9 shows four segments including TP, FP and FN examples. Here, the effect of post processing on the raw WaveNet model predictions can be observed.

In experiment 2, the multiclass model resulted in an overall accuracy of 97%. Mean performance metrics over the four respiratory event classes can be observed in Table IV. Performances vary



(a) Receiver operating characteristics curve

(b) Precision-recall curve

Fig. 3. ROC and PR curves for binary classification in experiment 1.

TABLE III
OVERALL PER-EVENT PERFORMANCE FOR EXPERIMENT 1 WITH ALL VALUES IN MEAN PERCENTAGES WITH 95% CONFIDENCE INTERVALS, AND EXPERIMENT 2 WITH MEAN PERFORMANCE AMONG THE INDIVIDUAL EVENTS, I.E. OBSTRUCTIVE APNEA, CENTRAL APNEA, RERA, HYPOPNEA

	Experiment 1		Experiment 2
	MGH dataset (binary)	SHHS dataset (binary)	MGH dataset (multiclass)
Accuracy	95.7 [95.7-95.7]	94.0 [94.0-94.1]	99.1
Sensitivity	67.7 [67.6-67.8]	70.9 [70.7-71.0]	49.3
Specificity	97.6 [97.6-97.6]	95.1 [95.1-95.2]	99.5
Precision	65.4 [65.3-65.5]	40.7 [40.5-40.8]	37.4
F1-score	66.5 [66.4-66.6]	51.7 [51.6-51.8]	40.6
Cohen's kappa	64.2 [64.1-64.3]	48.7 [48.6-48.9]	36.5 [36.4-36.6]

Note that for experiment 2 all performance metrics (except Cohen's kappa) do not show a 95% confidence interval range since these are mean values from all individual event types as seen in Table IV.

TABLE IV
PER-EVENT PERFORMANCE IN EXPERIMENT 2 INCLUDING MEAN PERCENTAGES AND 95% CONFIDENCE INTERVALS

	Sensitivity	Specificity
Obstructive apnea	50.6 [50.4-50.8]	99.7 [99.7-99.7]
Central apnea	84.1 [83.9-84.3]	99.6 [99.6-99.6]
RERA	39.5 [39.4-39.7]	99.0 [99.0-99.0]
Hypopnea	22.8 [22.7-23.0]	99.7 [99.7-99.7]
Mean	49.3	99.5
	Precision	F1-score
Obstructive apnea	44.5 [44.3-44.7]	47.4 [47.2-47.5]
Central apnea	41.5 [41.3-41.7]	55.6 [55.4-55.8]
RERA	24.9 [24.8-25.0]	30.6 [30.4-30.7]
Hypopnea	38.8 [38.6-39.0]	28.8 [28.6-28.9]
Mean	37.4	40.6

considerably for the different classes, e.g. while 84% of all expert-labeled central apnea events are correctly classified, this is only true for 23% of hypopneas. The absolute and normalized confusion matrices are shown in Appendix Table VI.

B. Per-patient performance

We next assessed the performance of our model to classify the severity of AHI. For the MGH dataset in experiment 1, performance among each AHI subgroup is shown in Table V. The sensitivity, precision and F1-score increased with the severity of apnea. The opposite effect was observed for the accuracy and specificity.

We computed the confusion matrix for AHI prediction, as shown in Fig. 4. Overall, 69% of patients from the MGH dataset and 54% from the SHHS dataset were assigned to the correct AHI category. Cohen's kappa values for AHI classification were 55% and 32% for the MGH and SHHS dataset, respectively. In experiment 2, 70% of all patients were classified in the correct AHI category using the MGH dataset. Cohen's kappa value was 56%. Most misclassifications resulted in false positives in neighboring AHI categories.

The scatter plots in Fig. 5 show the correlation between the expert-scored AHI and the model predicted AHI from experiment 1. The r^2 was 0.90 for the MGH dataset and 0.79 for the SHHS dataset. For experiment 2, an r^2 of 0.90, 0.84, 0.96, 0.96, 0.38 and 0.66 was determined for AHI, RDI, obstructive apneas, central apneas, RERAs and hypopneas, respectively, see Appendix Fig. 10.

The computed histograms represent the difference between the AHI value scored by the experts and the AHI value predicted by our model, see Fig. 6. The error distributions showed a clear peak around 0 with an AHI standard deviation of 7.8 and 7.4 for the MGH dataset in experiment 1 and experiment 2. For the AHI difference in the SHHS dataset the standard deviation was 9.5.

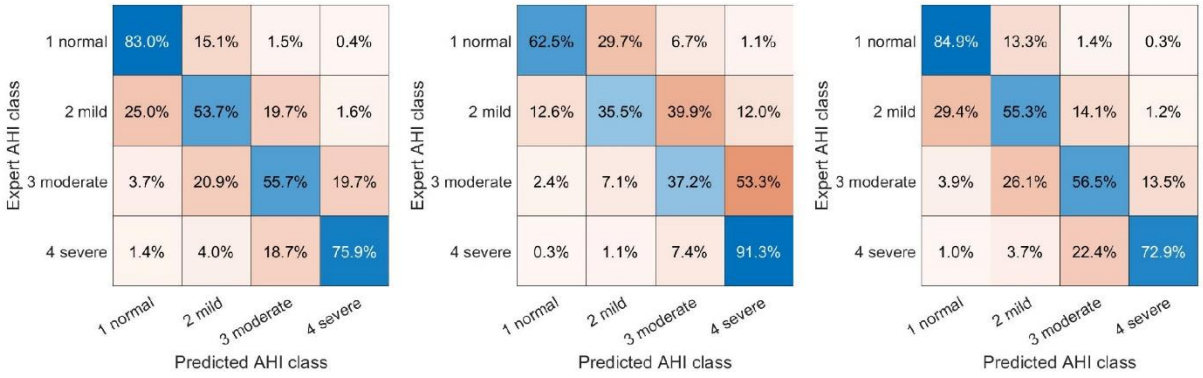
IV. DISCUSSION

A deep neural network method was developed to classify typical breathing disorders during sleep based on a single respiratory effort belt used in PSG. In a first experiment our WaveNet model successfully discriminated respiratory events from regular breathing

TABLE V

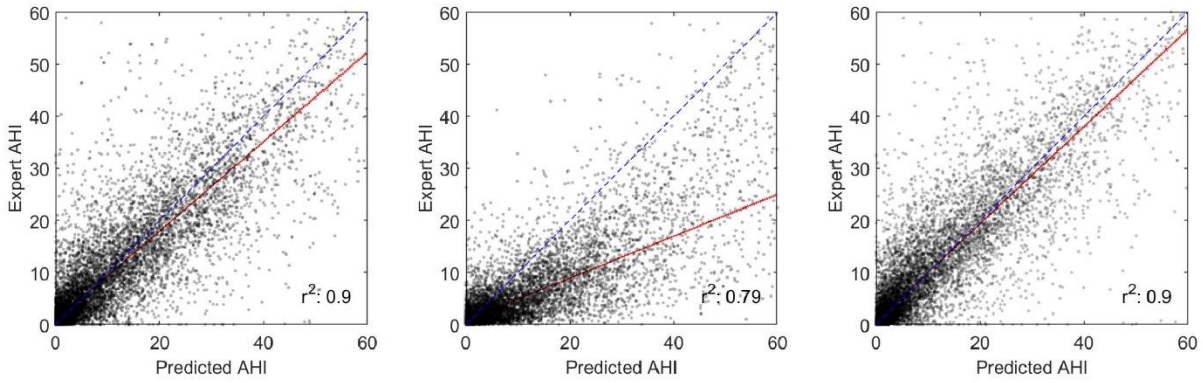
PER-EVENT PERFORMANCE PER AHI SUBGROUP FOR THE MGH DATASET IN EXPERIMENT 1 WITH ALL VALUES IN PERCENTAGES

	Accuracy	Sensitivity	Specificity	Precision	F1-score
Normal breathing (N=3660)	99	39	99	20	25
Mild apnea (N=2456)	96	46	98	33	36
Moderate apnea (N=1775)	94	51	96	39	42
Severe apnea (N=934)	91	50	95	43	44



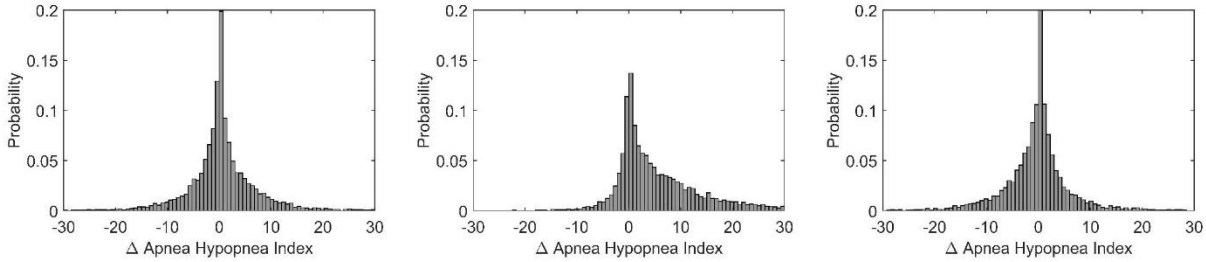
(a) MGH - binary classifier (b) SHHS - binary classifier (c) MGH - multi-class classifier

Fig. 4. AHI classification confusion matrices for both experiments with Cohen's kappa values of (a) 55%, (b) 32%, and (c) 56%.



(a) MGH - binary classifier (b) SHHS - binary classifier (c) MGH - multi-class classifier

Fig. 5. Scatter plots showing the correlation between the expert-scored AHI and the model predicted AHI in experiment 1 and 2. The fitted robust linear regression model is shown in red.



(a) MGH - binary classifier (b) SHHS - binary classifier (c) MGH - multi-class classifier

Fig. 6. Histograms of the relative difference in AHI determined by the experts vs the model in experiment 1 and 2.

on our primary dataset with an accuracy of 96%, and sensitivity, specificity, precision and F1-score of 68%, 98%, 65% and 67%, respectively. AHI was predicted for each patient using the number of respiratory events with an accuracy of 69%. It is notable that most misclassifications of our model resulted in false positives into the neighboring AHI categories. This effect is best visualized in the histograms in Fig. 6; the unimodal and symmetrical shape shows that a decrease in number of false positives was observed as the difference between the predicted AHI and the sleep-expert scored AHI increased. The correlation between expert-scored AHI and algorithm-predicted AHI showed an r^2 of 0.90. It is possible to adjust the predicted AHI cut -offs to improve AHI classification. However, we decided to use the original AASM criteria, because these are generally recognized as clinically meaningful and well understood categories.

When applying our model on a secondary dataset obtained from the SHHS, a slight decrease in model performance was observed. This is likely due to imperfect generalization to a dataset where different respiratory effort sensors are used. It is important to note that a thermistor was used to detect respiratory events in the SHHS study, whereas a nasal pressure sensor was used in the MGH study. There is agreement in the sleep field that nasal pressure-based scoring, regardless of ancillary signals used, is more sensitive in detecting sleep-disordered breathing than thermistor-based scoring [36]. Therefore, it is likely that a significant number of events were missed during annotation in the SHHS study. When observing the performance of our model applied on the SHHS dataset a decrease of approximately 15% was observed for the precision and f1-score. Sensitivity, however, slightly increased. This observation can be explained by the different methodology used while annotating events. The fact that our model generally overpredicts AHI when applied on the SHHS dataset, as seen in the confusion matrix of Fig 4 (b) and the scatter plot of Fig. 5 (b), are in line with this assumption. We recognize that the SHHS may have used ancillary signals. Nevertheless, use of the thermistor signal as the primary flow channel is one possible explanation for our results.

The guidelines for scoring respiratory events manually have evolved over the years but have remained largely driven by consensus. Thus, for example, the requirement of a 50% or 30% reduction in signal amplitude is arbitrary; there is no data to suggest that a 35% or 60% would be less or more clinically meaningful. Moreover, visual discrimination of small percentage differences is likely poor. During polysomnography or even home sleep study recordings, the multichannel nature of the data enables increased scoring accuracy by associating changes with neighboring signals. Moreover, airway collapse is common during central apnea, and high loop gain can drive obstructive events. Thus, the differentiation of obstructive and central events is not as pathophysiologically clear as clinical scoring may suggest. This biological reality of blurred boundaries will be reflected in any manual or automated scoring approach.

In a secondary experiment our model successfully identified the type of the included respiratory events, i.e. central apneas, obstructive apneas, hypopneas and RERAs. Despite a similar overall accuracy, discrimination of the specific respiratory events resulted in a decreased per-event performance with respect to the first experiment. Central apneas were detected with high sensitivity of 84%, expectedly due to the apparent effect of the disorder on respiratory effort. Often markedly reduced respiratory effort is observed during central apnea events, resulting in clear features for algorithms to recognize. We expect that this is the main reason of the high-performance metrics for the detection of central apneas. This is true to a lesser extent for obstructive apnea events, hence the slightly lower performance when compared to the central apneas. When using a single effort signal, thoracoabdominal asynchrony is undetectable. If using more than one effort sensor, this feature could enhance differentiation between obstructive apnea and central apnea by our model. The recognition of hypopneas and RERAs was considered moderate, with an F1-score of 31% and 29% respectively. The scatter plots show underprediction by our model, indicating limited sensitivity rather than specificity. Without additional information derived from other physiological signals the identification of hypopneas and RERAs appears difficult. It should be noted that scoring RERAs and central hypopneas are considered so difficult that the AASM scoring guidelines leaves these as "optional", and most clinical services do not score such events. There are also several biological inconsistencies with the conventional rules for scoring central hypopneas, adding to the probability of misclassification during "gold standard" scoring.

However, multiclassification often meant that an event of a particular respiratory class was classified as a different class. When observing per-patient performance of our multiclassification model, large variation in performance was observed among the various respiratory events. Yet, when the different predicted classes are grouped together to binary apnea events, a similar correlation was found between the expert-scored AHI and the algorithm-predicted AHI. An r^2 of 0.90 was determined, indicating that AHI prediction based on the specific respiratory events is feasible. Very similar performance was observed in AHI prediction confusion matrices with respect to the binary classification of experiment 1. The ability to discriminate various respiratory events is clinically valuable but may not be achievable by using manual scoring as a gold standard. The type of breathing assistance and overall apnea treatment may vary for different underlying pathology leading to apnea. Specifying the type of apnea will therefore provide aid in improving personalized patient care.

It is possible that valuable information was lost due to down sampling during preprocessing our data. However, the low pass filter of 10 Hz used for down sampling our signals was not expected to remove significant event characteristics that limit us in identifying apneas. Regular breathing for adults normally ranges between approximately 0.2 - 0.3 Hz.

Predicting for each second provides the smallest time resolution of our model. Reducing the rate at which the algorithm provides results can be achieved by aggregating predictions from consecutive time steps, such as taking the most severe form of respiratory event. This is one of the possible alternative approaches to generate events with a duration of 10 seconds or more and may yield a better performance.

Besides a high accuracy, a metric that is affected by class imbalance, our model also showed high AUC values for ROC (0.93), PR (0.71), and F1-score (0.67). This means the model not only has an excellent agreement in sensitivity and specificity but also has a clinically acceptable precision in specific situations, similar to the use of home sleep apnea testing, where tolerance to especially false negatives is required [37]–[39]. We have included the F1 score and the AUC_{PR}, as such performance metrics are not influenced by the imbalance of negative-positive classes but rather by sensitivity and precision of the positive class. The low standard deviation between the 5 folds of cross-validation (AUC_{ROC} and AUC_{PR} mean and std of 92 ± 0.5 and 71 ± 1.2) emphasizes the robustness of our model on a large dataset.

In manual analysis, experts learn to implicitly visually discount artifacts. Similarly, for the automated analysis, rather than designing algorithms by hand to explicitly address this issue, we took a data driven approach, i.e. presenting a large number of labeled examples including ones with artifacts present, and allowed the model to learn (implicitly) to discount artifacts. This is possible for two reasons: (1) The MGH dataset is very large; (2) deep neural network models,

like the one used in the manuscript, are very flexible. Thus, given sufficient data, deep neural network models can often learn to perform challenging pattern recognition tasks at a level that matches human experts. Our results show that this indeed was the case.

Most approaches found in the literature used different sensors to detect respiratory events. Some have shown slightly higher performances, although performance comparisons are difficult given the different datasets and evaluation methods. To our knowledge, our model showed better results with respect to other methods using a single respiratory effort belt and is the only model that shows that additional respiratory event class discrimination is possible based on respiratory effort only.

An advantage of using an effort belt to assess apnea is the non-invasive application. This becomes very relevant when assessing respiratory stability and instability/events in intensive care or environmentally hostile conditions. Using limited resources – such as a respiratory effort belt – to assess respiratory abnormalities can be successfully applied in combination with other simple and small sensors necessary for monitoring patients in diverse clinical situations. Patients receiving breathing aid using CPAP are eligible for event detection. The number of patients included in our research is larger than previous reports in the literature. This, in combination with limited preprocessing and without the use of any human-engineered features, emphasizes the robustness of our proposed approach.

V. Conclusion

A neural network approach to analyzing typical respiratory events during sleep based on a single respiratory measurement is described. Our model included dilated convolutions to allow their receptive fields to grow exponentially with depth, which is important to model the long-range temporal dependencies in respiration signals. Using this model, we obtained a comparable performance with respect to literature while using a minimally invasive methodology. Differentiation of event types is more difficult and may reflect in part the complexity of human respiratory output and some degree of arbitrariness in the clinical thresholds and criteria used during manual annotation. The use of a respiratory effort belt at the abdomen for sleep apnea analysis bears the advantage of wide implementation options ranging from acute care settings to wearable devices for home usage. Important first steps were obtained in automated apnea detection with limited resources, creating new sleep assessment opportunities applicable to the clinical setting.

References

[1] A. V. Benjafield, N. T. Ayas, P. R. Eastwood, R. Heinzer, M. S. Ip, M. J. Morrell, C. M. Nunez, S. R. Patel, T. Penzel, J. L. D. Pépin, P. E. Peppard, S. Sinha, S. Tufik, K. Valentine, and A. Malhotra, "Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis," *The Lancet Respiratory Medicine*, vol. 7, no. 8, pp. 687–698, Aug. 2019. DOI: 10.1016/S2213-2600(19)30198-5.

[2] T. L. Skaer and D. A. Sclar, "Economic implications of sleep disorders," *PharmacoEconomics*, vol. 28, no. 11, pp. 1015–1023, 2010. DOI: 10.2165/11537390-000000000-00000.

[3] J. B. Pietzsch, A. Garner, L. E. Cipriano, and J. H. Linehan, "An integrated health-economic analysis of diagnostic and therapeutic strategies in the treatment of moderate-to-severe obstructive sleep apnea," *Sleep*, vol. 34, no. 6, pp. 695–709, Jun. 2011. DOI: 10.5665/SLEEP.1030.

[4] M. S. Avidan, P. Strutz, W. Tzeng, B. Arrington, V. Kronzer, S. McKinnon, A. Ben Abdallah, and S. Haroutounian, "Obstructive sleep apnea as an independent predictor of postoperative delirium and pain: Protocol for an observational study of a surgical cohort [version 2]," *F1000Research*, vol. 7, 2018. DOI: 10.12688/f1000research.14061.2.

[5] S. L. Revels, B. H. Cameron, and R. B. Cameron, "Obstructive sleep apnea and perioperative delirium among thoracic surgery intensive care unit patients: Perspective on the STOP-BANG questionnaire and postoperative outcomes," *Journal of Thoracic Disease*, vol. 11, no. Suppl 9, S1292–S1295, 2019. DOI: 10.21037/jtd.2019.04.63.

[6] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of Big Data*, vol. 2, no. 1, p. 1, Dec. 2015. DOI: 10.1186/s40537-014-0007-7.

[7] L. Yue, D. Tian, W. Chen, X. Han, and M. Yin, "Deep learning for heterogeneous medical data analysis," *World Wide Web*, pp. 1–23, Mar. 2020. DOI: 10.1007/s11280-019-00764-z.

[8] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," Sep. 2016. arXiv: 1609.03499. [Online]. Available: http://arxiv.org/abs/1609.03499.

[9] M. Zabihi, A. B. Rad, S. Kiranyaz, S. Särkkä, and M. Gabbouj, "1D Convolutional Neural Network Models for Sleep Arousal Detection," Mar. 2019. arXiv: 1903.01552.

[10] S. Biswal, H. Sun, B. Goparaju, M. Brandon Westover, J. Sun, and M. T. Bianchi, "Expert-level sleep scoring with deep neural networks," *Journal of the American Medical Informatics Association*, vol. 25, no. 12, pp. 1643–1650, Dec. 2018. DOI: 10.1093/jamia/ocy131.

[11] T. Van Steenkiste, W. Groenendaal, D. Deschrijver, and T. Dhaene, "Automated Sleep Apnea Detection in Raw Respiratory Signals Using Long Short-Term Memory Neural Networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 6, pp. 2354–2364, Nov. 2019. DOI: 10.1109/JBHI.2018.2886064.

[12] R. Haidar, S. McCloskey, I. Koprinska, and B. Jeffries, "Convolutional Neural Networks on Multiple Respiratory Channels to Detect Hypopnea and Obstructive Apnea Events," in *Proceedings of the International Joint Conference on Neural Networks*, vol. 2018-July, Institute of Electrical and Electronics Engineers Inc., Oct. 2018, ISBN: 9781509060146. DOI: 10.1109/IJCNN.2018.8489248.

[13] S. McCloskey, R. Haidar, I. Koprinska, and B. Jeffries, "Detecting hypopnea and obstructive apnea events using convolutional neural networks on wavelet spectrograms of nasal airflow," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10937 LNAI, Springer Verlag, Jun. 2018, pp. 361–372, ISBN: 9783319930336. DOI: 10.1007/978-3-319-93034-3_29.

[14] N. Banluesombatkul, T. Rakthanmanon, and T. Wilaiprasitporn, "Single Channel ECG for Obstructive Sleep Apnea Severity Detection using a Deep Learning Approach," *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, vol. 2018-Octob, pp. 2011–2016, Aug. 2018. DOI: 10.1109/TENCON.2018.8650429. arXiv: 1808.10844.

[15] P. Lakhan, A. Ditthapron, N. Banluesombatkul, and T. Wilaiprasitporn, "Deep Neural Networks with Weighted Averaged Overnight Airflow Features for Sleep Apnea-Hypopnea Severity Classification," *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, vol. 2018-Octob, pp. 441–445, Aug. 2018. DOI: 10.1109/TENCON.2018.8650491. arXiv: 1808.10845.

[16] F. Espinoza-Cuadros, R. Fernández-Pozo, D. T. Toledano, J. D. Alcázar-Ramírez, E. López-Gonzalo, and L. A. Hernández-Gómez, "Speech Signal and Facial Image Processing for Obstructive Sleep Apnea Assessment," *Computational and Mathematical Methods in Medicine*, vol. 2015, 2015. DOI: 10.1155/2015/489761.

[17] G. Gutiérrez-Tobal, D. Álvarez, J. Gomez-Pilar, F. del Campo, and R. Hornero, "Assessment of Time and Frequency Domain Entropies to Detect Sleep Apnoea in Heart Rate Variability Recordings from Men and Women," *Entropy*, vol. 17, no. 1, pp. 123–141, Jan. 2015. DOI: 10.3390/e17010123.

[18] D. Álvarez, R. Hornero, D. Abásolo, F. Del Campo, and C. Zamarrón, "Nonlinear characteristics of blood oxygen saturation from nocturnal oximetry for obstructive sleep apnoea detection," *Physiological Measurement*, vol. 27, no. 4, pp. 399–412, Apr. 2006. DOI: 10.1088/0967-3334/27/4/006.

[19] T. Rosenwein, E. Dafna, A. Tarasiuk, and Y. Zigel, "Breath-by-breath detection of apneic events for OSA severity estimation using non-contact audio recordings," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2015-Novem, Institute of Electrical and Electronics Engineers Inc., Nov. 2015, pp. 7688–7691, ISBN: 9781424492718. DOI: 10.1109/EMBC.2015.7320173.

[20] S. H. Choi, H. Yoon, H. S. Kim, H. B. Kim, H. B. Kwon, S. M. Oh, Y. J. Lee, and K. S. Park, "Real-time apnea-hypopnea event detection

during sleep by convolutional neural networks," *Computers in Biology and Medicine*, vol. 100, pp. 123–131, Sep. 2018. DOI: 10.1016/j.compbiomed.2018.06.028.

[21] T. Kim, J. W. Kim, and K. Lee, "Detection of sleep disordered breathing severity using acoustic biomarker and machine learning techniques," *BioMedical Engineering Online*, vol. 17, no. 1, Feb. 2018. DOI: 10.1186/s12938-018-0448-x.

[22] R. Haidar, I. Koprinska, and B. Jeffries, "Sleep apnea event detection from nasal airflow using convolutional neural networks," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10638 LNCS, Springer Verlag, 2017, pp. 819–827, ISBN: 9783319701387. DOI: 10.1007/978-3-319-70139-4_83.

[23] H. ElMoaqet, J. Kim, D. Tilbury, S. K. Ramachandran, M. Ryalat, and C.-H. Chu, "Gaussian mixture models for detecting sleep apnea events using single oronasal airflow record," *Applied Sciences*, vol. 10, no. 21, p. 7889, Nov. 2020. DOI: 10.3390/app10217889. [Online]. Available: https://doi.org/10.3390/app10217889.

[24] S. S. Mostafa, F. Mendonça, A. G. Ravelo-García, and F. Morgado-Dias, *A systematic review of detecting sleep apnea using deep learning*, Nov. 2019. DOI: 10.3390/s19224934.

[25] 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," *Physiological Measurement*, vol. 39, no. 3, Mar. 2018. DOI: 10.1088/1361-6579/aaafb8.

[26] M. Bsoul, H. Minn, and L. Tamil, "Apnea MedAssist: Real-time sleep apnea monitor using single-lead ECG," *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 3, pp. 416–427, May 2011. DOI: 10.1109/TITB.2010.2087386.

[27] H. ElMoaqet, M. Eid, M. Glos, M. Ryalat, and T. Penzel, "Deep recurrent neural networks for automatic detection of sleep apnea from single channel respiration signals," *Sensors*, vol. 20, no. 18, p. 5037, Sep. 2020. DOI: 10.3390/s20185037. [Online]. Available: https://doi.org/10.3390/s20185037.

[28] T. Van Steenkiste, W. Groenendaal, J. Ruyssinck, P. Dreesen, S. Klerkx, C. Smeets, R. De Francisco, D. Deschrijver, and T. Dhaene, "Systematic Comparison of Respiratory Signals for the Automated Detection of Sleep Apnea," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2018-July, Institute of Electrical and Electronics Engineers Inc., Oct. 2018, pp. 449–452, ISBN: 9781538636466. DOI: 10.1109/EMBC.2018.8512307.

[29] D. of Sleep and C. Disorders, *Sleep heart health study*, 2019. [Online]. Available: https://sleepdata.org/datasets/shhs/variables/abdoqual.

[30] R. Yamamoto, *Wavenet vocoder*, 2019. [Online]. Available: https://www.github.com/r9y9/wavenet_vocoder.

[31] H. Sun, W. Ganglberger, E. Panneerselvam, M. J. Leone, S. A. Quadri, B. Goparaju, R. A. Tesh, O. Akeju, R. J. Thomas, and M. B. Westover, "Sleep staging from electrocardiography and respiration with deep learning," *Sleep*, vol. 43, no. 7, Dec. 2019. DOI: 10.1093/sleep/zsz306. [Online]. Available: https://doi.org/10.1093/sleep/zsz306.

[32] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, Oct. 2017. DOI: 10.1016/j.neunet.2018.07.011. arXiv: 1710.05381.

[33] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, p. 27, Dec. 2019. DOI: 10.1186/s40537-019-0192-5.

[34] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, Apr. 1960. DOI: 10.1177/001316446002000104. [Online]. Available: https://doi.org/10.1177/001316446002000104.

[35] MATLAB, *robustfit (R2020b)*. Natick, Massachusetts: The MathWorks Inc., 2020. [Online]. Available: https://www.mathworks.com/help/stats/robustfit.html.

[36] R. Budhiraja, J. L. Goodwin, S. Parthasarathy, and S. F. Quan, "Comparison of nasal pressure transducer and thermistor for detection of respiratory events during polysomnography in children," *Sleep*, vol. 28, no. 9, pp. 1117–1121, Sep. 2005. DOI: 10.1093/sleep/28.9.1117.

[37] J. A. Reichert, D. A. Bloch, E. Cundiff, and B. A. Votteri, "Comparison of the NovaSom QSG™, a new sleep apnea home-diagnostic system, and polysomnography," *Sleep Medicine*, vol. 4, no. 3, pp. 213–218, May 2003. DOI: 10.1016/s1389-9457(02)00234-4. [Online]. Available: https://doi.org/10.1016/s1389-9457(02)00234-4.

[38] N. Scalzitti, S. Hansen, S. Maturo, J. Lospinoso, and P. O'Connor, "Comparison of home sleep apnea testing versus laboratory polysomnography for the diagnosis of obstructive sleep apnea in children," *International Journal of Pediatric Otorhinolaryngology*, vol. 100, pp. 44–51, Sep. 2017. DOI: 10.1016/j.ijporl.2017.06.013. [Online]. Available: https://doi.org/10.1016/j.ijporl.2017.06.013.

[39] S. Su, F. M. Baroody, M. Kohrman, and D. Suskind, "A comparison of polysomnography and a portable home sleep study in the diagnosis of obstructive sleep apnea syndrome," *Otolaryngology–Head and Neck Surgery*, vol. 131, no. 6, pp. 844–850, Dec. 2004. DOI: 10.1016/j.otohns.2004.07.014. [Online]. Available: https://doi.org/10.1016/j.otohns.2004.07.014.

APPENDIX I
ADDITIONAL TABLES AND FIGURES

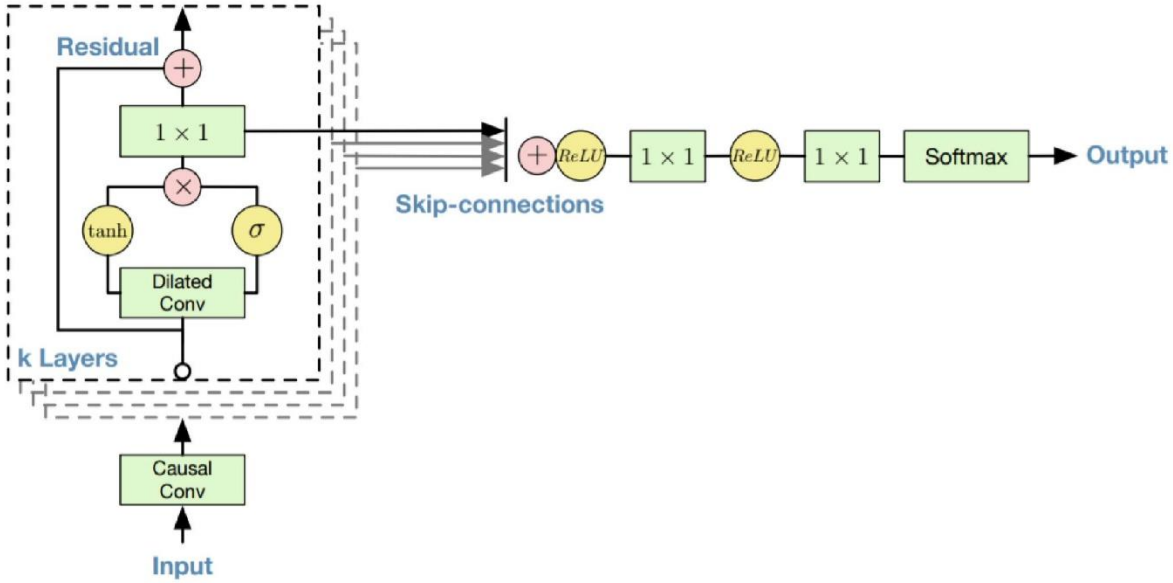


Fig. 7. Neural network architecture and residual block of the WaveNet model as described by Oord *et al* (2016).

TABLE VI
CONFUSION MATRICES FOR EXPERIMENTS 1 AND 2 IN BOTH ABSOLUTE AND RELATIVE VALUES.

Experiment 1, MGH absolute values	predicted No-event	predicted event
True, No-event	9694899	241920
True, event	218036	457631

Experiment 1, SHHS absolute values	predicted No-event	predicted event
True, No-event	7544601	385724
True, event	108492	264218

Experiment 1, MGH normalized values	predicted No-event	predicted event
True, No-event	0.98	0.02
True, event	0.32	0.68

Experiment 1, SHHS normalized values	predicted No-event	predicted event
True, No-event	0.95	0.05
True, event	0.29	0.71

Experiment 2, MGH absolute values	predicted No-event	predicted Obstructive	predicted Central	predicted RERA	predicted Hypopnea
True, No-event	36080590	114918	151308	378306	125924
True, Obstructive	28985	92177	22027	19704	18556
True, Central	10788	7090	107440	1548	1006
True, RERA	146761	13326	12413	125424	17857
True, Hypopnea	111845	60572	24466	72244	79817

Experiment 2, MGH normalized values	predicted No-event	predicted Obstructive	predicted Central	predicted RERA	predicted Hypopnea
True, No-event	0.97	0.0	0.01	0.02	0.0
True, Obstructive	0.16	0.51	0.12	0.11	0.10
True, Central	0.08	0.06	0.84	0.01	0.01
True, RERA	0.46	0.04	0.04	0.40	0.06
True, Hypopnea	0.32	0.17	0.07	0.21	0.23

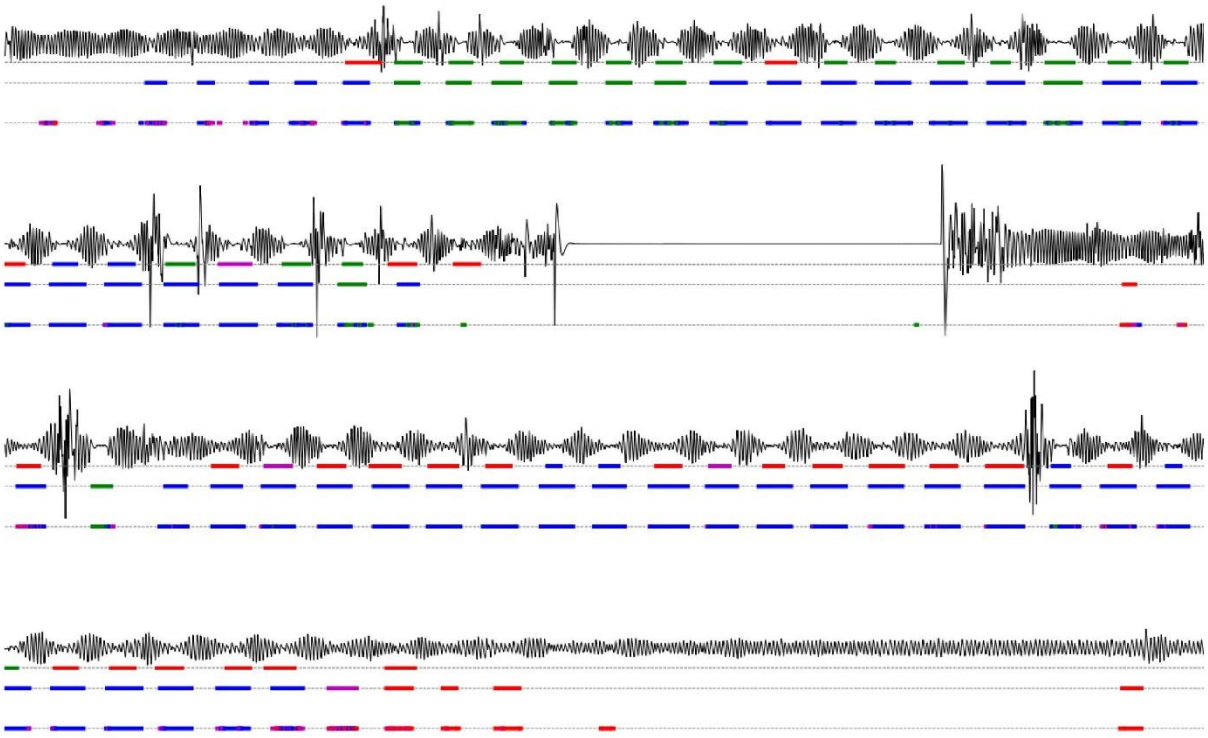


Fig. 8. Example recording showing the abdominal effort signal, the original expert labels, the smoothed WaveNet model output, and the raw WaveNet model output, from top to bottom. As the large flat parts show no continuous false positive apnea predictions we are convinced that our model learned not to classify such regions as apnea. With obstructive apneas in blue, central apneas in green, hypopneas in pink, and RERAs in red.

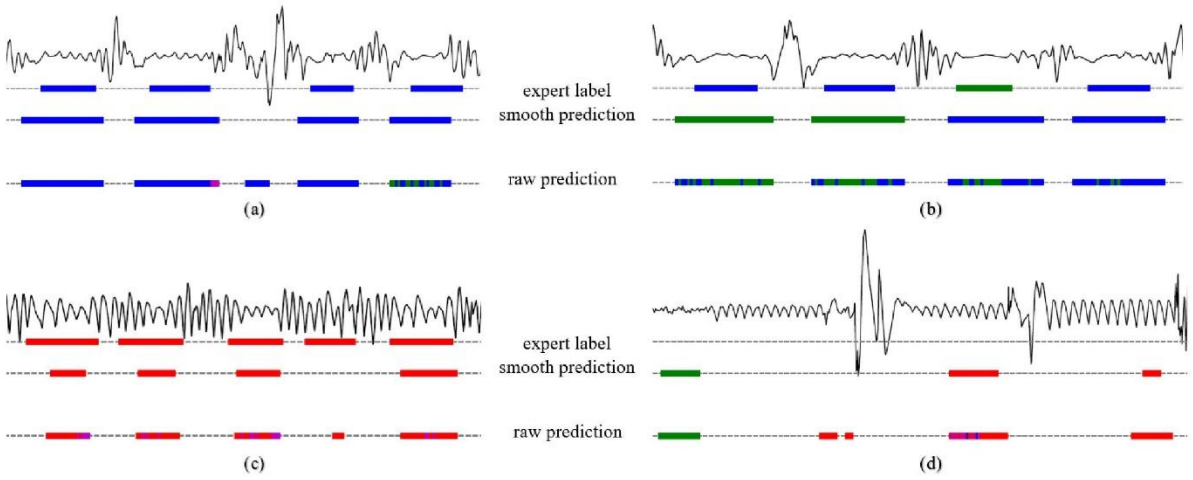


Fig. 9. Example signal segments and according labels and model predictions with in blue obstructive apneas, green central apneas, red RERAs and in pink hypopneas. (a), accurate predictions. (b), miss-classifications between obstructive and central apneas. (c), true positive and false negative RERA detections. (d), false positive event detections.

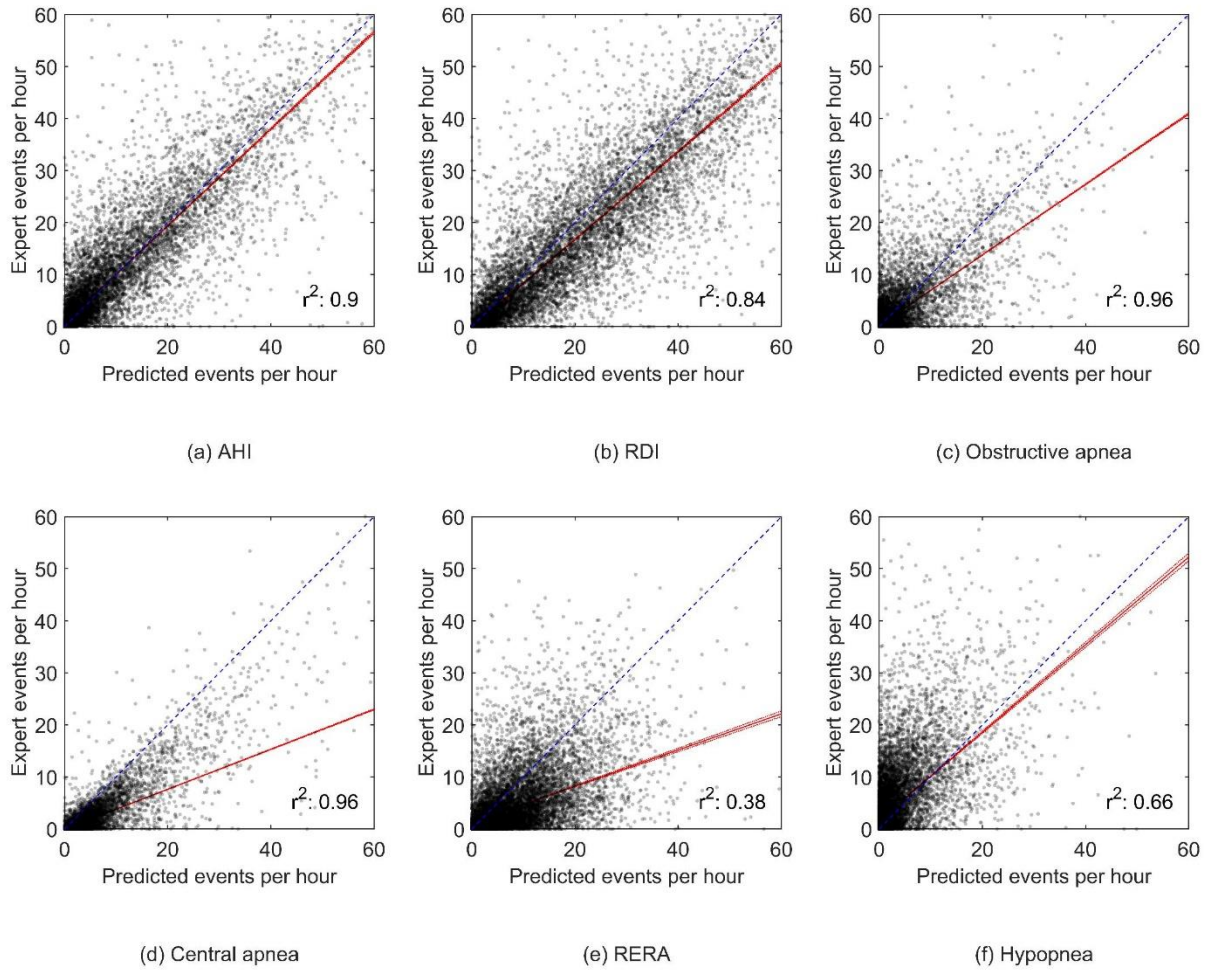


Fig. 10. Scatter plots showing the correlation between the expert-scored respiratory events and the model predicted respiratory events from experiment 2. The fitted robust linear regression model is shown in red.