Smart Health 27 (2023) 100373

Contents lists available at ScienceDirect

Smart Health

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

A clinical evaluation of a low-cost strain gauge respiration belt and machine learning to detect sleep apnea

Stein Kristiansen [a], Konstantinos Nikolaidis [a], Thomas Plagemann [a,*], Vera Goebel [a], Gunn Marit Traaen [b,a], Britt Øverland [c], Lars Akerøy [d,e], Tove-Elizabeth Hunt [b,a], Jan Pål Loennechen [e,d], Sigurd Loe Steinshamn [d,e], Christina Holt Bendz [b], Ole-Gunnar Anfinsen [b], Lars Gullestad [b,a], Harriet Akre [b,a]

[a] University of Oslo, Oslo, Norway
[b] Oslo University Hospital, Rikshospitalet, Oslo, Norway
[c] Lovisenberg Diakonale Hospital, Oslo, Norway
[d] St. Olavs University Hospital, Trondheim, Norway
[e] Norwegian University of Science and Technology, Trondheim, Norway

ARTICLE INFO

MSC:
41A05TODO
41A10TODO
65D05TODO
65D17TODO

Keywords:
Sleep apnea
Machine learning
Respiration belt
Strain gauge
Nox T3

ABSTRACT

Sleep apnea is a common and severe sleep-related respiratory disorder. Since the symptoms of sleep apnea are often ambiguous, it is difficult for a physician to decide whether to prescribe a clinical diagnosis, i.e., polysomnography (PSG), which results in a large percentage of undiagnosed and very late diagnosed cases. To reduce the time to diagnosis we investigate whether sleep monitoring data collected with a low-cost strain gauge respiration belt (called Flow) and a smartphone can be used to estimate with machine learning (ML) the severity of a patient's sleep apnea. The Flow belt and the Type III sleep monitor Nox T3 were used together by 29 patients for unattended sleep monitoring at home, resulting each in 235 hours of sleep data. Through experimental analysis, we found that Convolutional Neural Networks are best suited to analyze the Flow data, because they are most robust against the frequently occurring baseline issues and exhibit the best performance with an accuracy of 0.7609, sensitivity of 0.7833, and specificity of 0.7217. These results can be achieved even if the classifier is trained only on high-quality data from the Nox T3. Thus, there are good chances that future ML experiments with data from other low-cost respiration belts can benefit from existing open PSG datasets without new extensive data collection. On a low-end smartphone, the classifier needs approximately one second to analyze the sleep data from one night. The results demonstrate the potential of low-cost strain gauge belts, smartphones, and ML to enable large parts of the population to perform sleep apnea pre-screening at home.

1. Introduction

Sleep Apnea (SA) is a common and strongly under-diagnosed sleep-related respiratory disorder with severe health implications. It affects the natural breathing cycle during sleep and the disruption of normal airflow causes a decrease of oxygen in the blood. If the oxygen level is too low, the brain will force an awakening in order to resume normal breathing. The patient will most likely not remember the continuous awakenings. With such disrupted sleep, the person will in some cases never go into deep sleep, resulting

* Corresponding author.
E-mail address: plageman@ifi.uio.no (T. Plagemann).

https://doi.org/10.1016/j.smhl.2023.100373
Received 23 March 2021; Received in revised form 10 December 2022; Accepted 5 January 2023
Available online 11 January 2023

in daytime sleepiness and fatigue. When left untreated, SA can lead to serious medical conditions including cardiovascular diseases such as stroke, metabolic syndromes such as diabetes, and psychological disorders such as depression and anxiety (Huang, Qin, Zhang, & Chow, 2008; Punjabi, 2008; Young, Skatrud, & Peppard, 2004). In Norway, 16% of those aged 30–65 suffer from SA, and 8% are estimated to have moderate or severe SA (Hrubos-Strøm et al., 2011). Unfortunately, SA is often diagnosed very late or not at all. It is estimated that 70%–80% of those affected remain undiagnosed (Punjabi, 2008).

One factor contributing to SA being severely underdiagnosed is that feeling tired during the day is normal for many people. The patients typically have no recollection of the nightly awakenings. Therefore, patients can typically only describe vague symptoms to the physician that do not necessarily indicate the need for a clinical diagnosis with polysomnography (PSG) in a sleep laboratory (the gold standard for SA diagnosis) or polygraphy (PG) via unattended sleep monitoring at home.

Inspired by the popularity of health apps on smartphones and -watches, we investigate in our research the potential of wearables for low-cost SA pre-screening at home. The goal is to enable a broad population to independently perform sleep monitoring at home. Smartphones are a commodity and could be used for (1) data acquisition together with a wearable, (2) data analysis to detect SA events in the data recorded overnight and estimate the SA severity, and (3) carry the data and analysis results to a physician to give the physician a better foundation to prescribe the patient a clinical SA diagnosis. This could motivate patients to visit a physician and to undergo a clinical SA diagnosis before serious health implications are manifested. Such a solution needs to be (1) low-cost, (2) patient-friendly with few comfortable-to-wear sensors, (3) user-friendly with respect to the required computing infrastructure and its use, i.e., a single smartphone should be able to perform data acquisition and analysis, and (4) the monitoring data needs to be of sufficient quality such that machine learning (ML) can detect SA events with satisfactory performance.

In our earlier work, we could demonstrate that data from a single sensor from PSG and PG, like a respiration belt, can be used to detect SA events with an accuracy of 0.93 for PSG data (Kristiansen et al., 2018) and 0.83 for PG data (Kristiansen et al., 2020). Furthermore, we could demonstrate through controlled laboratory experiments that low-cost respiration belts (based on strain gauge technology) result in a data quality that comes close to the data quality of respiratory inductance plethysmography (RIP) belts used in PSG and PG (Løberg, Goebel, & Plagemann, 2018).

However, the difference between SA detection in PSG and PG data suggests that there will be a larger difference between the quality of monitoring data collected in controlled laboratory experiments and unattended sleep monitoring at home with a strain gauge respiration belt. The current state of knowledge is either based on studies with low-cost sensors used in a sleep laboratory or on clinically certified sensors used in unattended sleep monitoring at home. This work contributes to the existing knowledge with results and insights from a clinical study in an unattended home monitoring setting. We evaluate the ML performance with data from a low-cost strain gauge respiration belt in comparison to using data from the Nox T3 sleep monitor. The Nox T3 from Nox Medical (Reykjavik) is one of the most widely used portable Type III sleep monitor for SA diagnosis and records signals such as nasal airflow, respiratory effort with RIP belts at the chest and abdomen, pulse oximetry, activity, body position, and snoring.

In this study, we had the unique opportunity to test a strain gauge sensor, called Flow, with patients that underwent each two nights of PG with the Nox T3 (Nox, 2020). This was performed as part of a large clinical study, called the A3 study, at the Oslo University Hospital and St. Olavs Hospital, Trondheim (Traaen et al., 2020). A total of 579 patients were involved in the A3 study and the last 29 patients used the Flow sensor and the Nox T3 simultaneous for sleep monitoring. The majority of these patients recorded two nights of sleep data with the Flow sensor, resulting in recordings from a total of 49 nights. A sleep expert scored the Nox T3 data by labeling time periods that contain apneic events. We use these labels for the Flow data after synchronizing the Nox T3 and the Flow recordings. The Nox T3 and Flow data sets are used in this paper to investigate two basic research questions (RQ):

- RQ1: How accurately can ML models estimate SA severity with data from the Flow sensor?
- RQ2: Is it feasible to perform the data analysis on a smartphone with an acceptable response time?

This is to the best of our knowledge the first in-depth study of SA detection performance of ML with data from unattended sleep monitoring at home with a low-cost strain gauge sensor. Despite several issues with the Flow sensor, we achieve an accuracy of 0.7609, sensitivity of 0.7833, and specificity of 0.7217. Especially the high sensitivity of 0.7833 shows that such a low-cost sensor and ML can be used for SA pre-screening and provides valuable information to physicians that need to decide whether they should prescribe a SA diagnosis (with PSG or PG) for the patient. A detailed discussion leading to this conclusion is given in Section 5. With more elaborated pre-processing and a future version of the sensor that records the data on the sensor itself there is a high potential to improve these results. The analysis of a one-night sleep recording takes approximately one second on a low-end smartphone. We demonstrate that mixed training, i.e., using Flow data together with high-quality Nox T3 data to train a model for Flow data classification, positively impacts the achievable classification performance; and using only Nox T3 data as training data leads to comparatively good performance. This implies that training a good classifier for another strain gauge sensor might not require collecting a large amount of sleep data with that sensor.

The remainder of this paper is structured as follows: In Section 2 we present the background and related work. Section 3 describes the material and methods, including data acquisition, pre-processing, and experiments. The results of the experiments are presented in Section 4. In Section 5 we discuss the results, in Section 6 we elaborate the limitations of this work, and Section 7 concludes this paper.

2. Background and related work

This section gives a brief introduction to SA Diagnosis, the sensor technologies of the two respirations belts used in this study, and the existing body of work on ML for SA detection in general and ML with respiration belt data for SA detection in particular.

2.1. SA diagnosis

The gold standard for diagnosing SA is via PSG in a sleep laboratory (Punjabi, 2008). It requires the patient to stay overnight and record various physiological signals during sleep, such as electrocardiogram, electroencephalogram, electromyogram, electrooculogram, oxygen saturation, heart rate, blood pressure, and respiration from the abdomen, chest, and nose. A sleep expert is required to carefully analyze the data according to strict guidelines like those provided by the American Academy of Sleep Medicine (AASM) (Berry et al., 2012). This process involves identifying and denoting the beginning, end, and type of all periods with disrupted breathing. The two types of disrupted breathing are called apneas and hypopneas. Apneas are defined as a time period with a drop of airflow of at least 90% lasting for at least 10 s. Hypopneas are defined as a time period of at least 10 s with a drop of airflow of at least 30% breathing followed by a drop of oxygen saturation of at least 3%. A diagnosis is given based on the average number of such periods per hour, called the Apnea-Hypopnea Index (AHI). The AHI is used to classify the patient into one of four classes: (1) normal, no SA: AHI < 5, (2) mild SA: $5 \leq$ AHI < 15, (3) moderate SA: $15 \leq$ AHI < 30, and (4) severe SA: AHI ≥ 30.

PSG is very expensive, time-consuming, uncomfortable, and few sleep laboratories exist. A sleep expert is required to analyze and score the data. Type III and IV portable monitoring systems are intended to alleviate these problems and allow unattended sleep monitoring at home. Many PG devices are certified for clinical SA diagnosis (Collop et al., 2011; Mendonça, Mostafa, Ravelo-García, Morgado-Dias, & Penzel, 2018). In this study, one of the most widely used Type III monitor for PG, the Nox T3 from Nox Medical (Reykjavik, Iceland) is employed. The Nox T3 comprises a central recording unit with audio and accelerometer sensors, two RIP belts to measure movements of the chest and abdomen, nasal airflow sensor, and pulse oximetry. The Nox T3 is accompanied by software to download the monitoring data from the recording unit to a PC, perform sophisticated data pre-processing (including baseline corrections for the RIP belts), automatically analyze the data, and graphically present the monitoring data and analysis results. Furthermore, it allows exporting the monitoring data with a resolution of 1 Hz.

2.2. Respiration belt technologies

Respiration belts are used in PSG and PG to measure respiratory effort from the abdomen and chest. RIP is the gold standard (Berry et al., 2012), where coils are sewn into the flexible belts in a sinusoidal pattern and connected to an oscillator. Changes in the lengths of the belts during respiration alter the self-inductance of the coils which results in changes in the oscillations (Cohn et al., 1982). When properly calibrated, these changes accurately reflect the tidal volume of breathing (Chadha et al., 1982). The Flow belt used in our study is based on the much less expensive strain gauge technology (Sweetzpot, 2020). Variations in the lengths of the belts during breathing are measured as changes in the tension of the belt at a single point at the front of the chest or abdomen. While RIP belts measure actual changes in the belt length, strain gauges only estimate such changes at a single point and are as such susceptible to errors due to local variations in tension. Unlike the RIP technology, strain gauge has not been subject to rigorous clinical validation for sleep monitoring. The Flow sensor was originally intended for athletes to control breathing and is currently available for approximately 200 Euro. The RIP belts used for PG are an inherent part of the Nox T3. The necessary development efforts for hardware and software and the process of certification for clinical use lead to a substantially higher price for Type III sleep monitors compared to consumer electronics like the Flow sensor.

2.3. SA detection with ML

The achievable performance of automatic analysis of sleep data from clinically validated devices is well-known. There is a large body of research on automatic SA detection (Alvarez-Estevez & Moret-Bonillo, 2015; Faust, Acharya, Ng, & Fujita, 2016; Mendonca, Mostafa, Ravelo-García, Morgado-Dias, & Penzel, 2018; Mostafa, Mendonça, G Ravelo-García, & Morgado-Dias, 2019; Pombo, Garcia, & Bousson, 2017; Uddin, Chow, & Su, 2018). PG devices are intended for home monitoring (Collop et al., 2011), and most have support for (semi-) automatic analysis (Mendonça et al., 2018). They are however very expensive and normally use proprietary, custom-tailored analysis solutions. These solutions are not available for the growing variety of inexpensive sensors in the consumer market. For sensors of the latter kind, ML algorithms can be used. Our own and other works show that ML can accurately detect SA in PSG data (Kristiansen et al., 2018; Mostafa et al., 2019). We find that a single respiration belt at the abdomen performs almost as well as other sensors while being substantially more comfortable. We are aware of three works that use data from unattended home monitoring (Álvarez et al., 2020, 2016; Gutiérrez-Tobal, Álvarez, Crespo, del Campo, & Hornero, 2018), but these use expensive, clinically validated devices, and none investigate the performance obtainable using only a single respiration belt. Conversely, the works that do investigate ML performance with less expensive respiration belts do not use data from unattended home monitoring (Lin et al., 2016; Van Steenkiste et al., 2020). It is still an open research question what the impact of PSG in a sleep laboratory versus unattended sleep monitoring at home with PG or sleep monitoring with just one respiration belt is on sleep and sleep positions of patients (Mello et al., 2022) and the data quality. Therefore, it is important that devices that are intended for unattended sleep monitoring at home are evaluated in the same setting.

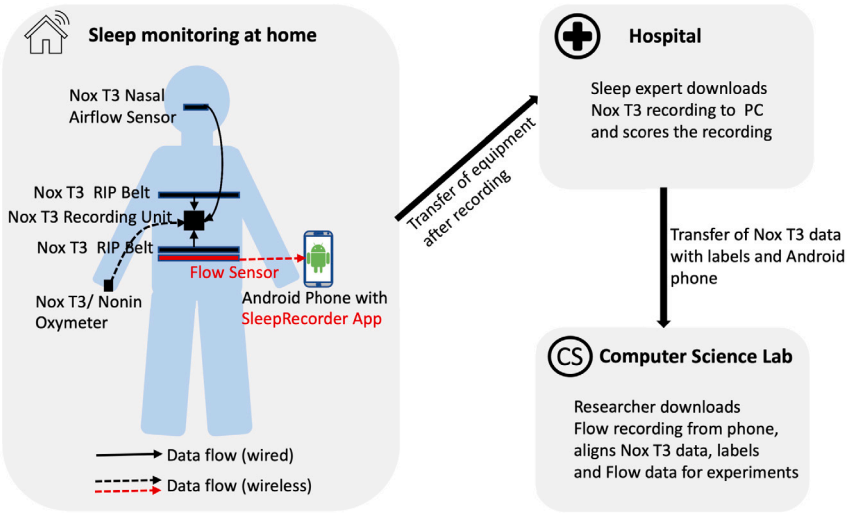


Fig. 1. From data acquisition to data sets. Sleep monitoring at home with Nox T3 for SA diagnosis in the A3 study (comprising the Nox T3 central recording unit, nasal airflow sensor, RIP belt on thorax and abdomen, and a Nonin oximeter on the fingertip) and the Flow sensor on abdomen (data recorded by the SleepRecorder App on an Android phone). Nox T3 data is labeled by a sleep expert in the hospital.

2.4. SA detection with respiration belt data and ML

The use of respiratory belts alone has not been equally well studied, especially with only one belt. We are only aware of six such works. In two works, disrupted breathing is simulated while awake (Dehkordi, Marzencki, Tavakolian, Kaminska, & Kaminska, 2012; Nepal, Biegeleisen, & Ning, 2002), which yields only approximate results. The other four works use data from sleeping patients in a controlled PSG setting (ElMoaqet, Eid, Glos, Ryalat, & Penzel, 2020; Lin et al., 2016; Tsouti, Kanaris, Tsoutis, & Chatzandroulis, 2020; Van Steenkiste et al., 2020), which is not ideal to study the performance of unattended home monitoring. Two of these works use RIP belts (ElMoaqet et al., 2020; Tsouti et al., 2020). A period-based accuracy of 84.4% was obtained with a single RIP belt on the abdomen (17 subjects) (ElMoaqet et al., 2020), and a subject-based accuracy 74% was achieved (12 subjects) (Tsouti et al., 2020). The other two works are based on less expensive piezo and bioimpedance technologies and are closer to our work (Lin et al., 2016; Van Steenkiste et al., 2020). However, they cannot be directly compared to our work, because the piezo-based respiration belt is part of the PSG system Alice 5 (Philips Respironics, Murrysville, PA) and Van Steenkiste et al. (2020) use the bio-impedance of the chest as a respiratory surrogate instead of respiratory effort that is measured by respiration belts. A period-based accuracy of 72.8% and 73.8% was achieved with bioimpedance (Van Steenkiste et al., 2020) (25 subjects) and piezo sensors (Lin et al., 2016) (34 subjects) at the abdomen, respectively.

3. Material and methods

This section describes the data acquisition, the pre-processing to handle data quality issues and prepare the data for the ML studies, and the experiments performed to investigate the research questions.

3.1. Data acquisition

In the A3 study, 579 subjects with paroxysmal atrial fibrillation were recruited to perform unattended sleep monitoring at home (most of them for two nights) with the Type III PG monitor Nox T3. Each individual was given brief training on sensor placement and use of the Nox T3 device before bringing the device home to undergo unattended sleep monitoring. We got the unique opportunity to additionally collect with the Flow sensor the respiration data from 29 patients. Fig. 1 shows which sensors these 29 patients used and the placement of these sensors. The patients were instructed on how to attach the Flow sensor side-by-side to the abdomen RIP belt of the Nox T3 and how to use the SleepRecorder app on an Android phone for data acquisition. Since both Flow and RIP belts are respiration belts and are placed at the same location, they measure the same physiological phenomenon, even if they are based on different technologies (see Section 2.2). The patients performed the sleep monitoring at home with both monitoring solutions simultaneously. All Nox T3 sensor data is stored on the central recording unit during unattended sleep monitoring at home. After the Nox T3 was returned to the hospital the data was downloaded from the central recording unit into the Noxturnal software on a PC. The Noxturnal software was used by a sleep expert to score the data. The recordings of Flow data were prone to temporary data loss due to Bluetooth disconnections and wrong timestamps assigned by the sleep recorder app. A landmark-based window approach was used to adjust the timestamps (Løberg et al., 2018). This way, the scoring from the sleep expert could be used to label periods of Flow data as either apneic or normal. Among the 29 individuals from which we collected data with the Flow sensor, 21 were male

Table 1
Size in terms of number of patients and hours of recording, and number of apneic and nonapneic periods in the data set from the A3 study (Nox) and the subset in which patient used Nox T3 and Flow in parallel (Flow).

Data set	# patients	Hours of recording	# apneic periods	# nonapneic periods
Nox	579	74 085	114 009	330 522
Flow	29	235	2988	10 117

Table 2
Data sets and their use in the experiments. All experiments use 10-fold cross validation except Experiment 3, which uses in each iteration the data of one patient as holdout test set. In all experiments, 30% of the training set is used as validation set.

Experiment	Training set	Test set
1 Impact of ML and BLA		
1.1 Flow (F)	F	F
1.2 Flow with BLA (FB)	FB	FB
1.3 Nox when Flow was used (NF)	NF	NF
2 Mixing high and low quality data		
2.1 Train on high quality test on low quality	N	FB
2.2 Sequential training on high and low quality	N → FB	FB
2.3 Train on high and low quality	NF ⋃ FB	FB
2.4 Train on all A3 data (N) and FB	N ⋃ FB	FB
3 Per subject based severity estimation	FB	FB
4 Inference on Smartphone	FB	FB

and 8 were female. The mean age was 61.1 (SD 10.9) and the mean BMI 28.2 (SD 4.8). The number in the SA severity classes mild, moderate, and severe is 13, 10, and one. Five individuals had an AHI below five, and thus belong to the class of normal individuals (no SA). A summary of the entire data from the A3 study and its subset in which the patient used the Nox T3 and Flow sensor in parallel is given in Table 1. The trial was performed according to the Declaration of Helsinki and approved by the Norwegian South-East Regional Ethics Committee (REK, ID: 2015/436) and the data protection officer at Oslo University Hospital. All patients provided written informed consent.

3.2. Pre-processing

A landmark-based window approach was used to adjust the wrong timestamps from the sleep recorder app and to detect missing data caused by Bluetooth problems (Andersen, 2020; Kristiansen et al., 2021). Movement during the night can result in short periods with motion artifacts in the sleep data which makes the corresponding data less useful for SA detection. This may subsequently result in a permanent change in the tension in the belt causing a sudden baseline shift in the recorded signal. The sensing technology in RIP belts is superior to strain gauge belts for such situations. The Flow sensor in addition suffers from gradual baseline wandering and drift. Nox Medical has spent a substantial effort to eliminate all baseline issues in the Nox T3 data. We design a simple Baseline Adjustment (BLA) procedure to reduce these issues for the data from the Flow sensor. BLA is achieved by standardizing each minute of the Flow data separately and repeating this process for all 1-minute windows over the entire recording (Andersen, 2020; Kristiansen, Andersen et al., 2021).

We refer to data from the Flow sensor as *Flow data*, and Flow data with and without BLA treatment as FB and F, respectively. Each of these comprises 235 h of data. Six additional recordings were present in the original data set, but had to be removed due to the complete absence of breathing signals and/or with only noise. We were unable to identify the reason for this signal corruption. A deep analysis of how this affects the quality of the recordings is found in Andersen (2020). We use two sub-sets of the data from the Nox T3 (*Nox data*). The first (N) contains the entire A3 data set from 579 individuals. Some recordings in the Nox data contain many artifacts. These artifacts are automatically detected by the software used to analyze and score the Nox data. We remove all recordings with more than 20% artifacts, leaving a total of 7408.85 h of data. The second sub-set, called NF, contains only the portion of the Nox data that was obtained at the same time as the Flow data (and was synchronized with it). This data set is of the same size as FB. Table 2 shows the data sets used for training and testing in our experiments. How they are used is described in the experiment descriptions below. The sampling rate of Flow and Nox T3 sensors is 10 Hz[1] and 100 Hz. Since breathing movements span several seconds, we can down-sample the data without loss of important respiratory changes. We down-sample the data to 1 Hz using the mean value of each second and standardize the resulting sample values. All data sets were finally standardized over the entire data set, such that the mean and variance of all values within one signal and one data set equals 0 and 1. Note that this is true also for FB, even if it was already subjected to a previous per-minute standardization with BLA.

[1] In practice, this sampling rate may vary slightly over time (Andersen, 2020).

3.3. Experiments

The main objective of this work is to understand and quantify the SA detection performance of ML with sleep monitoring data from a strain gauge respiration belt (RQ1) and whether the inference can be done on a smartphone (RQ2). To gain the necessary results and insights we design four experiments. In the following, we provide an overview of the experiments, and describe the procedures and metrics that are common to all experiments, followed by specific details of each experiment.

- Experiment 1: We study the classification performance of ML with the Flow data, and how the choice of ML technique and BLA impact the results. Thus, we train and test several state-of-the-art ML models on the Flow data without and with BLA.
- Experiment 2: We analyze the effect of augmenting the Flow data with the higher quality Nox T3 data for model training (mixed training) on classification performance. We use the best performing ML architecture from Experiment 1 (denoted CNNM) and train it on different combinations of Flow data and Nox T3 data.
- Experiment 3: We quantify how accurately CNNM can classify the SA severity of individual subjects. We use for each training and test iteration all data from one individual as test data. This data is excluded from the training data, which in turn is used as-is (imbalanced) and balanced.
- Experiment 4: We investigate whether the classifiers can in practice be used on a smartphone. We train CNN models on a powerful computer and use Tensorflow Lite to convert the classifier into a *float model* that is suitable to run on mobile devices. Individual one-night recordings are used to measure the inference time on a smartphone.

3.3.1. Common procedures and metrics

All experiments except Experiment 3 are conducted using 10-fold cross-validation (CV). It is common to randomly shuffle the data set before CV (called *pre-shuffling*). This is not done in our experiments, since it may result in unrealistically high classification performance in our envisioned scenario. We assume that a pre-trained classifier is used to estimate the SA severity of an individual from which data was not available during training. With pre-shuffling, training data will contain a significant amount of samples from the same individuals as those in the test set. In addition to the 10-fold CV we perform CV on a per-individual basis in Experiment 3. Since we have 29 individuals, this corresponds to 29-fold CV where the test-fold is always limited to the recordings from one given individual, and the training- and validation folds consist of recording from the remaining individuals. See Section 3.3.4 for further details.

We estimate SA severity via binary classification of periods. For comparability with related works, we classify each 60-second period as either *apneic* or *normal*.[2] Each data set with Flow data (F, FB, and NF) has 2988 apneic periods and 10 117 normal periods, and N has 114 009 apneic and 330 522 normal periods. Due to the imbalanced class distributions, we balance the data set using majority sub-sampling before dividing it into 10 folds, if not explicitly stated otherwise. We discard the final periods to ensure equal-sized folds.

Training is performed via mini-batch learning with a batch size of 1000 periods and a learning rate of 0.001. 30% of the training set is used as validation set. All classifiers were trained for 500 epochs. For all Neural Networks (NN), we use the cross-entropy loss function and the Adaptive Moment Estimation (ADAM) optimizer. In this paper, the choice of NN and their hyperparameters is based on our earlier work (Kristiansen, Nikolaidis et al., 2021) in which we conducted a systematic comparative study of classification performance and resource use with different combinations of 27 classifiers with the A3 data set, i.e., N.

We quantify classification performance using Cohens κ metric (Cohen, 1960). It is ideal for comparing classifiers as it accurately summarizes most aspects of prediction performance and accounts for random chance (Cohen, 1960). In addition, we present accuracy, sensitivity, and specificity since these have more intuitive semantics that can help understand the clinical relevance of the results. The presented results are those from the epoch with the highest κ on the validation set.

The experiments are conducted on a computer with a 56-core, 2.2 GHz, Intel Xeon Gold 5120 CPU, 128 GB RAM, and four Nvidia GTX 2080 Ti GPUs. In Experiment 4, we also use a Samsung Galaxy Tab S3. All software is implemented using Tensorflow v. 1.13.1.

3.3.2. Experiment 1: The impact of ML technique and BLA

We include a wide range of ML classifiers that differ in architecture, approach, size, and sophistication. Most are based on modern deep NN. For comparison, we include two traditional techniques that are commonly used in related works, i.e., Random Forest (RF) and Multi-Layer Perceptrons (MLP). The MLP are feed-forward NN with one hidden layer, and all layers are densely connected. The deep NN architectures include CNN, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). Since we use time-series data, for which LSTM are designed, we include several improvements and modifications of LSTM. These include BI-directional LSTM (BILSTM), Stacked BILSTM (SBILSTM), and SBILSTM With Attention (BIWALSTM) (Wang, Huang, Zhu, & Zhao, 2016). All classifiers are instantiated in the three sizes small, medium, and large. Classifiers are referenced with the acronyms of their architecture with a post-fix -S, -M, and -L for small, medium, and large-sized classifiers, respectively. The hyperparameter values for each classifier are presented in Table 3. We use the same values as in Kristiansen, Nikolaidis et al. (2021), which were tuned using N. We avoid tuning the hyperparameters using the Flow data to avoid biasing their values to our test sets. No additional feature extraction is used before training and testing, i.e., the 60 values in the 60-second periods mentioned above are used directly as input to all classifiers.

[2] We do not distinguish between obstructive, central, or mixed apneas and hypopneas in this work, as it does not affect the AHI, which is the main metric used to score SA severity.

Table 3
Classifier parametrization based on Kristiansen, Nikolaidis et al. (2021).

Classifier	Hyperparameter	Small	Medium	Large
MLP *Hidden layers: 1*	Hidden nodes	50	100	200
CNN *Kernel size: 5*	Conv. layers Filters/conv. layer Dense layers x nodes	3 32–128 1 × 128	4 32–256 1 × 256	5 32-512 2 × 256
LSTM	Hidden state size	50	100	200
GRU	Hidden state size	50	100	200
BILSTM	Hidden state size	50	100	200
BIWALSTM	Hidden state size	50	100	200
SBILSTM *Hidden state: 100*	Layers	1	3	5
RF *Max. nodes/tree: 100*	Trees	50	200	500

We perform three sub-experiments: one without BLA using F (Experiment 1.1), one with BLA using FB (Experiment 1.2), and as benchmark one with NF (Experiment 1.3). In each experiment, the different folds from the same data set are used for training and testing via 10-fold CV.

3.3.3. Experiment 2: Mixing high- and low-quality data

We perform four sub-experiments and repeat them 10 times. Due to these repetitions, the total number of experiments is very large. Therefore, we focus on the potential improvements beyond what was achieved in Experiment 1.2 and we use the CNNM classifier that achieves the overall best performance. In Experiment 2.1, we train the classifier on N and test it on FB. This reflects a scenario where a classifier trained on high-quality data is made available and used *as-is* to classify low-quality data. The remaining experiments use a mix of Flow (FB) and Nox data (N) for training. In Experiment 2.2, the training is performed *sequentially*, i.e., first with data from the Nox T3 then with data from the Flow sensor. A complete 10-fold CV training procedure is first completed with N. The resulting pre-trained models (one per fold) are subsequently used as the initial models for a complete 10-fold CV run using FB as both training and test data. The results are relevant for scenarios where a classifier trained on high-quality, closed data from, e.g., clinical studies is made available for further training on available low-quality data. In Experiments 2.3 and 2.4, training is performed *simultaneously* with Nox T3 and Flow data. In Experiment 2.3, each of the training sets produced during 10-fold CV is concatenated with NF, and only FB is used for testing. This configuration is found in scenarios where both high- and low-quality training data is available from the same set of individuals. Experiment 2.4 is identical to 2.3, except NF is replaced with N. This gives an estimate of performance obtainable in scenarios where high-quality data is available from many more individuals than for low-quality data. To study the difference between training with a mix of high- and low-quality data and training with either only high- or only low-quality data, we compare results from the above experiments with results obtained by repeating Experiments 1.2 and 1.3 10 times.

In Experiments 2.2, 2.3., and 2.4, Flow and Nox data are weighted differently during training by multiplying the loss of samples in the Nox data by $w \in [0, 1.0]$ before they are used for back-propagation. We execute these sub-experiments multiple times with varying w to estimate the optimal weight and to study the relative impact of low- and high-quality data during training.

3.3.4. Experiment 3: Per-subject SA severity estimation

It is only possible to estimate the SA severity of an individual by analyzing the entire data set from this individual. Therefore, we change the CV procedure that is used in Experiments 1 and 2. For each iteration in the CV procedure, the test set consists of all the recordings of a given subject in FB, while the training set consists of the remaining recordings in FB. To study the effect of balancing the training set, we compare results with classifiers trained on balanced (C_b) and imbalanced data (C_i). In addition, we create the combined classifier C_c by averaging the predictions from C_b and C_i. These classifiers use the best-performing CNNM architecture in Experiment 1.

A proper understanding of classification performance with imbalanced data requires a joint assessment of several metrics (He & Garcia, 2009). Since Experiment 3 focuses on SA-severity estimation, most of the results are based on comparisons between *predicted* and *actual AHI*. We compute actual AHI by counting the actual number of breathing events that are scored by a sleep expert as apneas or hypopneas in the Nox data.[3] Predicted AHI is computed based on the total number of predicted apneic periods. Since the number of apneas/hypopneas may differ from the number of apneic periods, the results in Experiment 3 reflect the negative performance impact of dividing the data into periods. This is not the case in most related works since online data sets often only include per-period labels (Mostafa et al., 2019). As a result, Experiment 3 yields a lower, but more realistic performance.

[3] Note that this AHI differs from the actual AHI of the subjects presented in Section 3.1, since a significant portion of the Nox-data is removed during pre-processing (see Section 3.2).

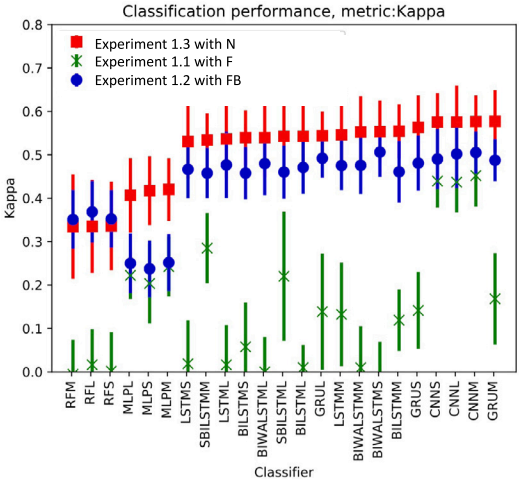
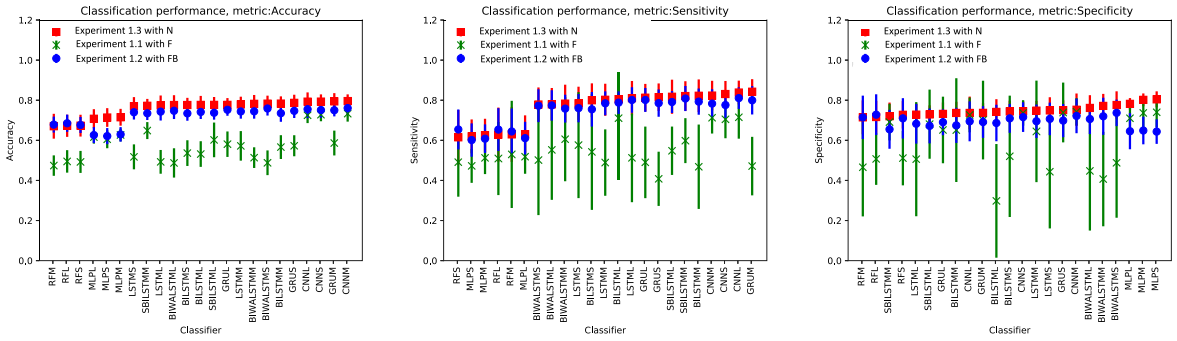


Fig. 2. κ for Experiment 1. BLA improved the performance of all classifiers, for most with large gains.



(a) (b) (c)

Fig. 3. Accuracy, sensitivity, and specificity for Experiment 1. Sensitivity is slightly higher than specificity.

3.3.5. Experiment 4: Inference on a smartphone

In order to perform the data analysis on a smartphone, we construct float models with Tensorflow Lite. These models are smaller than the original models without a significant loss in prediction performance and can be executed using the low-overhead *tflite* operations in Tensorflow Lite. The model can be *quantized* to reduce inference times further but at the cost of numerical precision and classification performance. We measure the inference times of both Float and quantized versions of the CNN models from Experiment 1. We measure the time spent classifying data from one complete night (from FB), as this reflects the response time of the data analysis the morning after a night of data collection. To study the impact of model size, we compare results from CNNS, CNNM, and CNNL. After converting the trained models, they are used for inference on a Samsung Galaxy Tab S3 with a 1.6 GHz quad-core Qualcomm Snapdragon 820 processor and 4 GB of RAM, running the Android 7.0 operating system and using Tensorflow version 2.2.0. Since the main focus of Experiment 4 is inference efficiency and the impact of model reduction on performance, 10 training epochs are sufficient to obtain representative results.

4. Results

4.1. Experiment 1: The impact of ML technique and BLA

The results of Experiment 1 in terms of κ are presented in Fig. 2, and the accuracy, sensitivity, and specificity in Fig. 3. Overall, sensitivity is slightly higher than specificity, indicating that the classifiers are better at correctly classifying apneic than non-apneic periods. Deep NN-based classifiers substantially outperform the older RF and MLP classifiers. The highest κ of 0.5062 $\pm$ 0.0575 is obtained with BIWALSTMS and BLA, closely followed by CNNM with a κ of 0.5052 $\pm$ 0.0485. The highest accuracy is obtained with CNNM and BLA, i.e., 0.7609 $\pm$ 0.0284, with a sensitivity of 0.7833 and a specificity of 0.7217. Surprisingly, RF yields better results with Flow data than with Nox data. However, RF yields the overall lowest performance for all metrics and data sets, probably because RF fails to properly learn the important features of the Nox data.

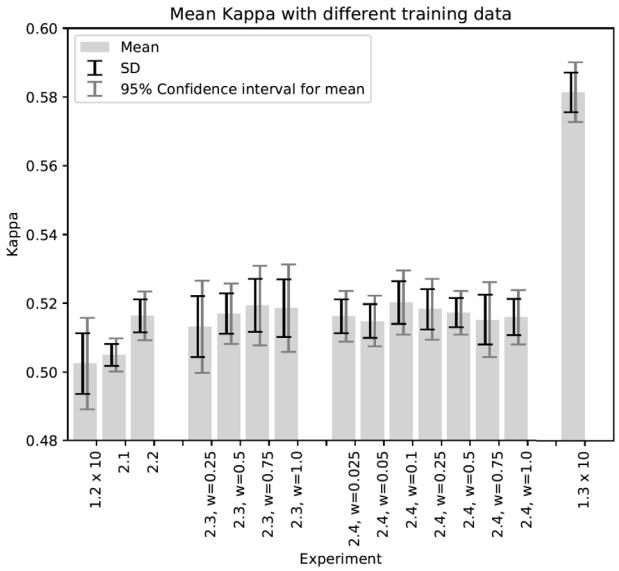


Fig. 4. Results from Experiment 2. The label for each column identifies the sub-experiment and the value of w (if used). w is the weight by which the loss from samples in Nox data is multiplied before back-propagation. Classifiers trained on N classify FB with high κ and the highest performance is achieved when using a mix of N and FB for training.

BLA clearly increases accuracy. On average, BLA leads to an increase in κ of 0.2889 ± 0.1091 points for any given classifier. BLA also has an equalizing effect on sensitivity and specificity, which is most prominent for MLP and GRU classifiers. CNN classifiers benefit far less from BLA than the others, with a mean increase in κ of only 0.1263 ± 0.1054. This indicates that CNN is resilient to baseline shifts and wandering.

CNNM and BIWALSTM obtain nearly the same performance, but require very different amounts of time for training and testing. While BIWALSTM requires 224 ms and 18 ms on average per epoch for training and testing, respectively, CNNM only requires only 18 ms and 2 ms. While such differences can be ignored when training on powerful GPUs and a sample rate of only 1 Hz, they become important when considering to deploy the models on resource-constrained devices like smartphones, especially with sensors that have much higher sampling frequencies (see Section 4.4).

4.2. Experiment 2: Mixing high- and low-quality data

The results from Experiment 2 are presented in Fig. 4. The x-axis shows the different sub-experiments, and the y-axis the mean κ for CNNM. This is not the mean across folds, as in Experiment 1. It is the mean across the results from 10 different repetitions of the complete 10-fold CV procedure with different random seeds. The result from each such repetition is the mean across the 10 folds. This makes visible the variance resulting from different initializations of the weights in the NN. When the training set consists of both Nox and Flow data, the weight on the Nox data is denoted as w in the labels on the x-axis.

We obtain similar κ in Experiment 2.1 (using only Nox data for training) as in Experiment 1.2 (using only Flow data for training). In both cases, performance is evaluated on the Flow data as the test set. This result suggests that classifiers that are pre-trained on high-quality data can be used *as-is* to classify data from different sensors of lower quality, and still obtain performance comparable to that of training on data from the target, low-quality sensor. Our hypothesis to explain the fact that a model trained only on Nox data (N) can generalize well enough to classify Flow data (FB) is based on the signal quality of FB compared to N (Kristiansen, Andersen et al., 2021). The mean sensitivity per recording in terms of the amount of measured breaths ranges from 97.2% to 89.5%. The mean positive predictive value (PPV) ranges from 94.2% to 82.1% and the mean weighted average percentage error (WAPE) is 18.4%. That means breaths can be detected quite well, but there are some errors in the amplitude accuracy. Therefore, we hypothesize that the core features for humans to detect apnea events are with some errors also captured by the Flow sensor, and a model that is trained on these features in N can detect them in FB and predict apnea events as presented in our results.

The results show an overall improvement in classification performance when using a mix of Nox and Flow data for training. The κ from Experiments 2.2–2.4 are between 0.0107 and 0.0177 higher than in Experiment 1.2. While the mean accuracy in Experiment 1.2 is 0.7580, the mean accuracies in Experiments 2.2–2.4 are 0.7650, 0.7647, and 0.7647, respectively. There is no clear difference in κ among Experiments 2.2–2.4 since differences in the mean values are much smaller than the corresponding SD. This suggests that it does not matter much for the performance: (1) If we first train on Nox data then on Flow data (Experiment 2.1), or train on both in parallel (Experiments 2.2–2.4). (2) If we use the complete A3 data set (Experiment 2.4) or only the sub-set obtained simultaneously with the Flow data (Experiment 2.3).shortlabel (3) How the Nox and Flow data is weighted during training.

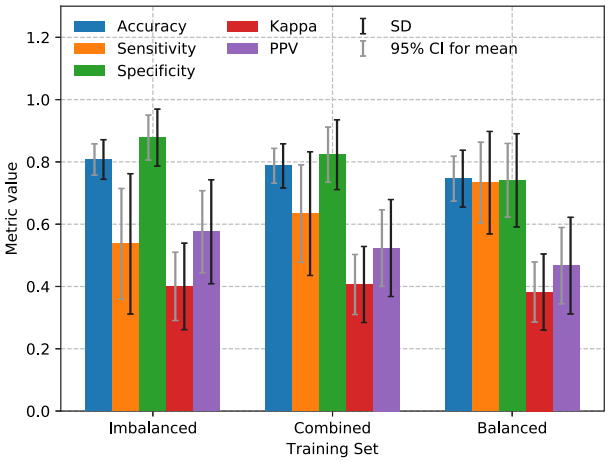


Fig. 5. Results from Experiment 3. The choice of using the original imbalanced, combined, or balanced training set has the most impact on sensitivity and specificity.

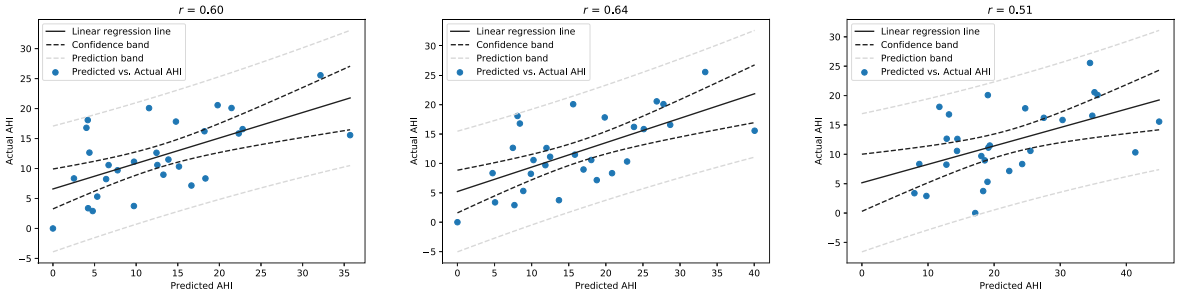
In Experiment 2.3, the training benefits from the preliminary synchronization of Nox and Flow data. In Experiment 2.4, training benefits from the large amount of training data available. The similarity of the results from these two experiments suggests that these two factors have a similar positive impact. This means that using the complete, large-scale Nox T3 data set from the A3 study yields a similar performance as using the much smaller sub-set that was collected at the same time as the Flow data. Due to the high SD and 95% CI we cannot rule out the existence of small differences between these two approaches that would be uncovered with a larger data set.

To properly compare the approaches we must also take into account the training times. The training data used in Experiment 2.3 is only twice as big as that used in Experiment 1.2. The mean training time per epoch is only about twice as long (71 ms). The data set used in Experiments 2.1, 2.2, and 2.4 comprises the complete A3 data set for training, which comprises 232 batches compared to only four batches in Experiment 1.2. The mean training times are significantly larger (1708 ms per epoch). While Experiments 1.2 and 2.3 were completed in minutes, Experiment 2.1, 2.2, and 2.3 require several hours to complete. This large difference in training time implies that it might be preferable to use NF instead of N to augment the training data from F in situations requiring significant amounts of training. Our study was not representative for such a situation mainly because the CSV data has a very low sampling rate of 1 Hz imposed by the Noxturnal software used for CSV export. However, it is not uncommon with much higher sampling rates. As an example, the original A3 Nox and Flow data were captured at 100 and 10 Hz, respectively, before being downsampled to 1 Hz.

4.3. Experiment 3: Per-subject SA severity estimation

As a basis to understand the accuracy of estimating the AHI of individuals, we first investigate how accurately the 60-second periods are classified for each individual. The mean, SD, and CI across all individuals are shown in Fig. 5. We see that the imbalanced classifier C_i outperforms the balanced classifier C_b for all metrics except sensitivity. The combined classifier C_c performs somewhere in between C_i and C_b. These findings are a result of the imbalanced test set, which has an over-representation of normal periods. This problem of *intrinsic imbalance* is common in the biomedical domain (He & Garcia, 2009), and the smaller the data set is, the larger its impact (Weiss & Provost, 2001). This configuration leads to different biases in the training set which is learned by the classifier: a balanced training set has a very different class distribution than the test set, while an imbalanced training set contains many more normal than apneic periods. C_i benefits in terms of accuracy and specificity from the much higher number of non-apneic (10 117) than apneic (2988) periods. C_b has seen an equal number of apneic and non-apneic periods, and the training set of C_i contains many more non-apneic ones. Therefore, C_i classifies more accurately non-apneic periods, which are much more numerous in the test set than apneic ones. This explanation is confirmed by the fact that C_b outperforms C_i on the balanced validation set in terms of all metrics except accuracy and specificity. These results should be seen in the light of the relatively large CI, indicating that a larger data set is desirable to increase the confidence in the mean values.

All remaining results are based on comparisons between the actual AHI of an individual and the AHI estimated using the classified periods for that individual. Fig. 6 shows scatter plots that compare actual and predicted AHI per individual. These include the 95% confidence and prediction bands that outline the 95% confidence intervals for the mean and individual AHI for the ground truth given predicted AHI, respectively. Fig. 6(a–c) show that we obtain r-values between 0.51 and 0.64 with different prediction approaches. Interestingly, C_c outperforms C_b and C_i, in terms of the r-value and the magnitude of the prediction and confidence intervals. There is a clear correlation between predicted and actual AHI. As expected, the r-values are lower than what can be obtained with manual scoring of data from clinically certified PG equipment. For instance, in Chang et al. (2019) an r-value of 0.791 is obtained when comparing manual scoring of Nox T3 and PSG data. This is explained by the results in Fig. 5. Values for



(a) C_i. (b) C_c. (c) C_b.

Fig. 6. Scatter plots for Experiment 3. The results are with C_i, C_c, and C_b from left to right, respectively. C_c outperforms C_i and C_b in terms of r-value and the magnitude of prediction and confidence intervals.

Table 4
Results for Experiment 3. Performance of binary classification with thresholds corresponding to the different SA severity classes. NaN: Not a Number (incomputable).

Training set	Threshold	Acc.	Sens.	Spec.	κ	PPV	P(A)	P(A|P)
Imbalanced (C_i)	5	0.8276	0.8400	0.7500	0.4487	0.9545	0.8621	0.9545
	15	0.7586	0.6364	0.8334	0.4781	0.7000	0.3793	0.7000
	30	0.9310	NaN	0.9310	0.0000	0.0000	0.0000	NaN
Balanced (C_b)	5	0.8621	1.0	0.0	0.0	0.8621	0.8621	0.9259
	15	0.5517	0.8182	0.3889	0.1786	0.4500	0.3793	0.6000
	30	0.7586	NaN	0.7586	0.0000	0.0000	0.0000	NaN
Combined (C_c)	5	0.8621	0.9600	0.2500	0.2658	0.8889	0.8621	0.8276
	15	0.7241	0.8182	0.6667	0.4528	0.6000	0.3793	0.4500
	30	0.9310	NaN	0.9310	0.0000	0.0000	0.0000	NaN

sensitivity show that only 53.3% to 73.3% of the apneic periods for each subject are correctly identified as such. We see that the slope of the regression line is lower than 1, which means that predicted AHI are more dispersed than the actual AHI. The fact that the performance in our results is not as good as previous works with clinically certified equipment is not surprising. A major reason is the much lower quality of data from consumer electronics like the Flow sensor. While recent deep NN are known to excel in classification even of noisy data, the presence of irreducible errors in both data and labels imposes an upper limit on the achievable classification performance. This limit is studied in literature both theoretically and empirically (Hestness et al., 2017; Tumer & Ghosh, 1996). Furthermore, a large amount of data is required to achieve this upper-performance limit (Kristiansen, Nikolaidis et al., 2021). Intuitively, the amount of data required increases with decreased data quality. The Flow data is of lower quality than the Nox data and much smaller. With just 29 subjects, personal differences have a non-neglectable impact on the ability of the classifiers to generalize.

Table 4 shows the classification performance where subjects are divided into two groups, one with AHI above and one below the thresholds (TH) that separate severity groups. Since no subjects belong to the actual severity group *severe* SA, the corresponding sensitivity, κ, and PPV values are either not computable or not meaningful. The $P(A)$ column presents the prior probability of a positive, based on the SA prevalence among the 29 subjects; and the $P(A|P)$ column presents the posterior probability of an actual positive given a positive test as given by Bayes' formula $P(A|P) = \frac{P(P|A)P(A)}{P(P)}$. The difference between the two right-most provides an estimate of how much a positive test increases the probability of an actual positive that accounts for SA prevalence (as opposed to PPV). Table 4 shows that C_b over-estimates SA severity. Only half of the subjects are correctly classified at $TH = 15$, and only 38.89% of subjects with AHI < 15 are classified as such. C_c yields good results for $TH = 15$, which is reflected in (1) a relatively high accuracy for $TH = 15$, (2) the same sensitivity as with C_b, and (3) a relatively modest decrease in specificity compared to C_i. We see that C_i outperforms the others for all thresholds in terms of accuracy, specificity, κ, and PPV, with the exception of $TH = 5$ for which C_c achieves slightly higher accuracy. C_b and C_c outperform C_i in terms of sensitivity. $TH = 15$ yields the lowest accuracy but the highest κ among all TH. This is explained by the relatively balanced class distribution at this threshold, i.e., 62.07% of the subjects have an actual AHI above this threshold. With $TH = 5$, 86.2% of the subjects are above the threshold and no one is above the threshold with $TH = 30$. Such imbalances result in artificially high accuracy. The usefulness of a test depends on how much a positive test result increases the probability an actual positive has (relative to the prior probability). The columns $P(A)$ and $P(A|P)$ show that C_i leads to the largest increase for all TH. Importantly, with $TH = 15$, a positive test result almost doubles the probability of SA from 37.93% to 70.00%.

4.4. Experiment 4: Inference on a smartphone

Table 5 presents the results for Experiment 4. *Original* refers to the classifier variant used in Experiments 1–3. Column *Inf. time* presents the mean time spent for inference per 60-second period of data in milliseconds (ms). Columns κ and *Acc.* present

Table 5
Results from Experiment 4. Mean inference time is the mean time spent for inference per period, i.e., the total amount of time spent for 480 periods divided by 480.

Classifier	*Parameters*	*Size*	*Inf. time*	κ	*Acc.*
CNNS:					
Original	166.658	2–2.2 MB		0.4903	0.7512
Float	166.658	654 kB	0.55 ms	0.4989	0.7525
Quantized	166.658	168 kB	0.34 ms	0.4928	0.7492
CNNM:					
Original	413.058	5–5.2 MB		0.5052	0.7609
Float	413.058	1617 kB	1.09 ms	0.4820	0.7458
Quantized	413.058	411 kB	0.55 ms	0.4882	0.7492
CNNL:					
Original	1.069.186	12.8–13.1 MB		0.5020	0.7554
Float	1.069.186	4181 kB	1.93 ms	0.4674	0.7375
Quantized	1.069.186	1055 kB	0.93 ms	0.4531	0.7308

classification performance in terms of the mean κ and accuracy across the 10 folds. Classification performance for the original models is obtained from Experiment 1.2, and therefore shows the mean across the 10 models from 10-fold CV. The float and quantized models are based on the first model from 10-fold CV. Since we only execute models intended for mobile devices on the portable device (float and quantized variants), Table 5 does not include inference times for the original models. Instead, we present the size of the original model for comparison. This size depends on the method used to store the model. We present the two sizes resulting from (1) just saving the weights as a checkpoint, in which case we need to specify the architecture of the model whenever we load it, and (2) saving also the architecture of the model.

Analysis of data from one complete night is in most cases completed in less than a second. This is true even for the float models. This result is obtained with a relatively low sample rate of 1 Hz. Sampling frequencies between 1 Hz and 100 Hz is not uncommon for portable sensors. Higher sampling frequencies generate larger data sets, which require more time for data analysis. With our data set, one night of data requires far less than 1 MB of working memory, even if we would include up to four simultaneous signals. With today's smartphones, memory and computational power will not constitute a bottleneck even at higher sampling frequencies.

We observe almost no reduction in prediction performance of float models compared to original models. For CNNS, the mean performance of the float model is slightly higher than that of the original model. This difference is most likely attributed to the high SD in our results, as seen in Fig. 4. The low inference times demonstrate that the classification performance obtained in Experiments 1–3 can be achieved on smartphones. We observe almost no performance reduction for quantized models, even if they reduce inference time by 38.18–51.81% compared to the float models. This can significantly reduce inference time with larger models or with sensors with higher sampling rates.

5. Discussion

Our results are similar to existing works with low-cost respiratory belts based on other technologies like piezo or bioimpedance technology. For instance, a maximum accuracy of 0.728 and 0.7380 is obtained in Lin et al. (2016) and Van Steenkiste et al. (2020). As expected, these results are not as good as one can achieve with data from PSG or PG, mainly because of the much lower quality of the inexpensive sensors used to collect the data in this study. This quality is certainly not sufficient for clinical diagnosis, but is it good enough for future pre-screening solutions? An indicator of sufficient quality is the agreement of human scorers; if the agreement between the presented ML-based low-cost solution and the human scores on the Nox T3 data is close to the agreement between human scorers performing SA diagnosis, then the results should be good enough for SA pre-screening. In the literature, the agreement between different scorers ranges from 76% mean agreement rate (Van Steenkiste et al., 2020) to an intraclass correlation coefficient of 0.95 (Magalang et al., 2013). Lin et al. (2016) question whether it makes sense to aim for higher accuracy than 80% due to the fact that the agreement of human scorers is around 80% according to their insights. Since it is our aim to reduce with SA pre-screening the number of undiagnosed patients, sensitivity is a very important metric, which is 0.7833 in our study. Another metric that demonstrates that automated analysis of Flow data yields predictions that correlate well with actual SA severity, is the correlation coefficient r. We achieved r-values of up to 0.64, which shows that the presented approach can be used to indicate SA severity. The performance in this study is limited mainly by the relatively simple and inexpensive construction of the Flow sensor, but also by Bluetooth, timestamp issues, and the simplicity of BLA. To reveal the full potential of the strain gauge technology for SA pre-screening, we suggest for future studies to timestamp the samples on the sensor itself, add storing capacity on the sensor to avoid data loss during Bluetooth disconnection, and develop more elaborated BLA preprocessing. We furthermore show that our classifiers can be used on small, resource-constrained devices, with almost no loss in classification performance. The CNN can analyze the data from a complete night mostly within one second on a smartphone from 2012 (Samsung Galaxy Tab S3). Modern smartphones, especially those with hardware support for ML-based inference, are expected to perform this task substantially faster. That means no additional computing infrastructure is needed for data acquisition and analysis.

The results from Experiment 2 suggest that classifiers that are pre-trained on high-quality data can be used *as-is* to classify data from different sensors of lower quality, and still obtain performance comparable to that of training on data from these low-quality

sensors. This has significant implications since the collection of large amounts of data is expensive and may not be possible for many producers of inexpensive home electronics due to privacy and resource restrictions. If one set of classifiers trained on high-quality data can be used as-is to analyze data from low-cost sensors, the cost of collecting large amounts of sensitive data can be substantially reduced, greatly facilitating low-cost screening based on consumer electronics. If the high-quality data set is not large enough, the results suggest that this may in some cases be compensated by collecting a much smaller amount of data from high- and low-quality sensors at the same time. We show that this results in substantial savings of training time compared to training with the complete, large-scale data set. This finding strongly motivates future work to evaluate the performance of the pre-trained classifiers on test data from additional low-cost sensors, in order to investigate the degree to which the same pre-trained classifiers can be used for a variety of low-cost sensors.

Another finding is that classifiers based on CNN outperform the others in several respects. First, they rank among the best-performing classifiers in terms of classification performance. Second, they appear to be the most resilient to baseline shift and wandering. Third, they require an order of magnitude less computational resources than RNN. The fact that it performs almost as well with and without BLA distinguishes it from the other ML techniques, and shows that it is by far the most robust ML technique against baseline shifts and wandering. This is not an obvious outcome, since CNN is not designed particularly for time-series data like RNN. To the best of our knowledge, our results are the first to demonstrate that this applies also to clinical data from consumer electronics. These results emphasize the versatility of CNN and its usefulness even for physiological time-series data.

Careful attention should be given to the decision of whether or not to balance the training data. This decision must be properly aligned with the objectives and available resources for SA severity estimation. Such estimation precludes balancing the test set, causing the class distribution of the test set to be determined by the prevalence of apneic periods. A key insight from Experiment 3 is that the class balance of the *training set* has a significant effect on SA severity estimation, and that this effect varies with the prevalence of SA in the data set. Balanced training data leads to fewer false negatives which is generally preferable from a purely medical perspective. But it leads to a consistent over-estimation of SA severity, which may be too expensive from a large-scale, societal perspective, because it may lead to a large number of unnecessary medical examinations. Conversely, imbalanced training data results in overall higher prediction accuracy, but at the cost of a higher number of false negatives, which has obvious negative implications. Imbalanced data yields the highest increase in the Bayesian posterior, i.e., the probability of actual SA upon a positive test result. This is often regarded as a key indicator of the usefulness of a medical test, which is noteworthy since applied ML studies commonly use balanced training data. We show that a compromise can be made by averaging the predictions from balanced and imbalanced training data. Interestingly, averaged predictions are better than those from balanced training data for discriminating between subjects with AHI above or below 15, which commonly determines whether to prescribe PSG. This approach slightly outperforms the use of balanced or imbalanced training data in two more ways, i.e., in terms of the r-value and the width of prediction and confidence bands (Fig. 6).

6. Limitations of the study

Most data sets used in this study are relatively small, both in terms of the amount of data and in terms of their sample frequency. Most of the evaluated classifiers are based on deep learning that benefits from large data sizes, while F, FB, and NF include relatively few and short recordings. The benefit of a large number of recordings is demonstrated in our previous work where we show that classification performance increases when including up to 3–400 recordings, where nearly all recordings contain data from two full nights (Kristiansen, Nikolaidis et al., 2021). This is at least one order of magnitude more data than in F, FB, and NF. Since Flow and Nox data are of the same type, it would be interesting to study the performance achievable with many more recordings. Although Nox and Flow data are collected at 100 and 10 Hz, an upper sampling rate of 1 Hz is imposed on the final data by the software used to export Nox data and the need to synchronize Nox and Flow data. As argued in Kristiansen et al. (2018), 1 Hz well exceeds the frequency of respiration and thus is sufficient to capture respiratory changes. However, this does not preclude the existence of higher frequency features in the breathing signal that can help distinguish between normal from disrupted breathing. It would therefore be interesting to study performance also with higher sampling frequencies.

We include only the Flow low-cost sensor in our study. One important finding is that a CNN classifier pre-trained on Nox data can be used without further training to classify Flow data with a relatively high classification performance. If this is the case also with other, similar low-cost sensors, it would be possible to use the same pre-trained classifiers for a wide range of low-cost sensors and thus significantly reduce the cost of data collection. To substantiate this claim, it is necessary in future work to reproduce this result with other low-cost sensors.

The ground truth in our study is obtained using PG, not PSG which is widely regarded as the gold standard for SA diagnosis. PG also involves manual scoring by a qualified sleep expert, but it is nevertheless a clinically validated alternative for SA diagnosis and is used in previous clinical studies such as Traaen et al. (2020). As such it constitutes a valid ground truth for our ML studies. The main drawback of PG compared to PSG is that is does not include equally many sensors. Due to the lack of electroencephalograph (EEG) sensors, for instance, it is not possible to perform arousal-based scoring, which is recently advised by the AASM to be part of the diagnosis of obstructive SA (Malhotra et al., 2018). Without an EEG sensor, it is for instance not possible to detect hypopneas associated with EEG-based arousal. Thus, a more reliable ground truth could be achieved with PSG. On the other hand, a key advantage of the PG data used in our study is that it is obtained through unattended sleep monitoring at home which is far more representative for the target scenario (patients perform pre-screening at home) than with PSG data collected in a sleep laboratory.

Six recordings were removed from the Flow-data after they were manually inspected and found to be corrupt. We cannot expect such manual inspection to occur in a practical screening scenario. Such corrupt recordings were however clearly distinguishable from non-corrupt recordings via visible inspection, as they exhibit large periods with nearly no visible breathing patterns. We are therefore convinced that such recordings can be identified automatically by anomaly detection algorithms that look for either manually or learned features in the recordings.

7. Conclusions and future work

The results of this work show that despite the limitations of the simple strain gauge technology, including frequent baseline issues, low-cost respiration belts like the Flow belt can be used to collect valuable sleep monitoring data such that modern ML is able to estimate the SA severity. Such a solution is neither intended nor able to replace existing methods for SA diagnosis, like PSG and PG. Instead, it can enable large portions of the population to perform pre-screening at home, because the costs for such respiration belts are low, smartphones are for many individuals a commodity, the belt is easy to attach and comfortable to wear, and the ML performance to estimate SA severity is close to what ML can achieve with PG data. Another use-case for this solution is the long-term monitoring of patients to observe how patients respond to treatments.

Our results have useful implications for future SA research and development. The fact that a model trained only on high-quality Nox T3 data can classify Flow data with high performance simplifies future studies with new respiration belts since extensive data collection with the new sensor is not necessary. Sleep monitoring data is like many other biomedical data sets rather imbalanced. The decision of whether the training data should be balanced or not has important implications on different aspects of the performance. Balanced training data leads to fewer false negatives and over-estimated SA severity, while imbalanced data results in overall higher prediction accuracy at the cost of more false negatives. Balanced training data may be a more appropriate alternative when sufficient resources are available to handle a large number of false positives, and imbalanced training data for large-scale deployment where such a cost is too high. We propose a third alternative that combines predictions from both approaches to yield a compromise.

We aim in future work to confirm the results and improve the performance. To substantiate the finding that data from various inexpensive sensors can be analyzed by the same classifier trained on data from high-quality sensors, it is necessary to study additional low-cost sensors. More advanced baseline correction algorithms might improve the data quality and increase classification performance. Improvements in hardware are expected to improve performance since we use sensors of low quality which impacts performance. Our current data set is relatively small, especially for the low-quality data. We expect the performance to improve with more training data. Finally, we intend to investigate personalized training and leverage data from subjects that are made available at an early stage.

CRediT authorship contribution statement

Stein Kristiansen: Conceptualization, Methodology, Software, Investigation, Data curation, Writing – original draft. **Konstantinos Nikolaidis:** Software, Investigation, Writing – review & editing. **Thomas Plagemann:** Conceptualization, Resources, Data curation, Writing – review & editing, Funding acquisition. **Vera Goebel:** Conceptualization, Resources, Writing – review & editing, Funding acquisition, Project administration. **Gunn Marit Traaen:** Data curation, Writing – review & editing. **Britt Øverland:** Data curation, Writing – review & editing. **Lars Akerøy:** Data curation, Writing – review & editing. **Tove-Elizabeth Hunt:** Data curation, Writing – review & editing. **Jan Pål Loennechen:** Data curation, Writing – review & editing. **Sigurd Loe Steinshamn:** Data curation, Writing – review & editing. **Christina Holt Bendz:** Data curation, Writing – review & editing. **Ole-Gunnar Anfinsen:** Data curation, Writing – review & editing. **Lars Gullestad:** Data curation, Writing – review & editing, Funding. **Harriet Akre:** Data curation, Writing – review & editing.

Declaration of competing interest

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

Data availability

The authors do not have permission to share data.

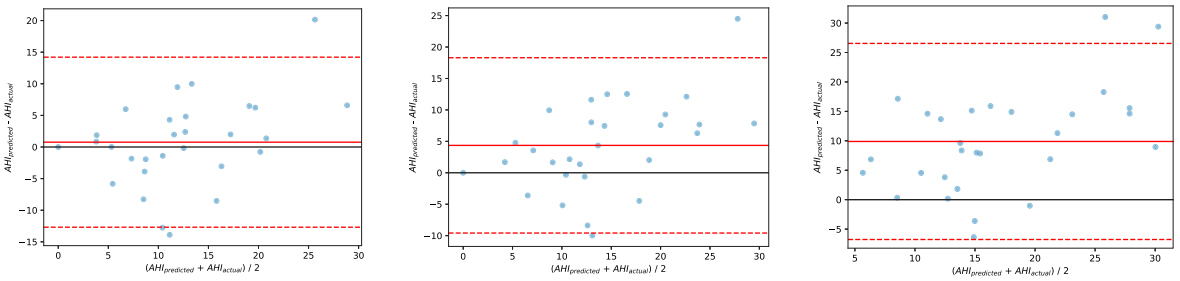
Acknowledgments

This work has been funded by the Research Council of Norway, Grant/Award Number: 250239/O70 (Cesar project), the Oslo University Hospital, Norway, the University of Oslo, Norway, and the Norwegian Health Association, Norway.

Appendix

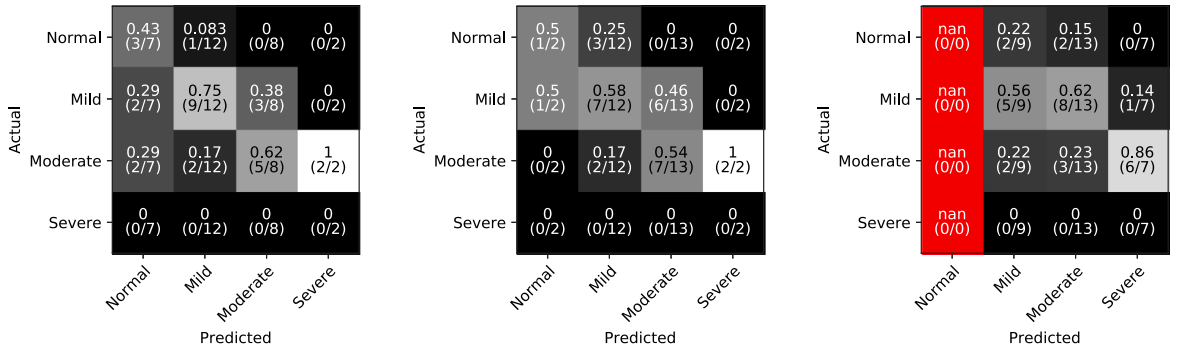
This appendix includes supplementary analysis results for the per-subject SA severity estimation in Experiment 3. Fig. 7 shows Bland-Altman plots that compare actual and predicted AHI per individual. Fig. 8 shows the confusion matrices for SA severity groups. They denote the probability of a subject belonging to a certain actual severity group given a predicted group. According to the balanced predictions, no subjects belong to the severity group *mild*, i.e., these values cannot be computed. The actual number of subjects in severity groups normal, mild, moderate, and severe is four, 14, 11, and zero, respectively.

The Bland-Altman plots show that C_b is biased towards higher SA severity groups, which is not the case for C_i. The effect is reflected in the confusion matrices. While C_i most often correctly predicts the SA severity group, C_b tends to overestimate SA severity. C_b has a higher sensitivity and a lower PPV than C_i (see Fig. 5), causing C_b to classify more periods as apneic on average.



(a) C_i. (b) C_c. (c) C_b.

Fig. 7. Bland-Altman plots for Experiment 3. The results are with C_i, C_c, and C_b from left to right. C_b is biased towards higher SA severity groups, which is not the case for C_i.



(a) C_i. (b) C_c. (c) C_b.

Fig. 8. Confusion matrices for Experiment 3 with the fraction of individuals in predicted SA severity groups that were actually in that severity group. C_i predicts most often correctly the SA severity group, while C_b shows a tendency to overestimate SA severity.

References

Álvarez, D., Cerezo-Hernández, A., Crespo, A., Gutiérrez-Tobal, G. C., Vaquerizo-Villar, F., Barroso-García, V., et al. (2020). A machine learning-based test for adult sleep apnoea screening at home using oximetry and airflow. *Scientific Reports*, *10*(1), 1–12.

Álvarez, D., Gutiérrez-Tobal, G. C., Vaquerizo-Villar, F., Barroso-García, V., Crespo, A., Arroyo, C., et al. (2016). Automated analysis of unattended portable oximetry by means of Bayesian neural networks to assist in the diagnosis of sleep apnea. In *2016 global medical engineering physics exchanges/pan american health care exchanges (GMEPE/PAHCE)* (pp. 1–4). IEEE.

Alvarez-Estevez, D., & Moret-Bonillo, V. (2015). Computer-assisted diagnosis of the sleep apnea-hypopnea syndrome: a review. *Sleep Disorders*, *2015*.

Andersen, M. H. (2020). *Analyzing the usefulness of a low-cost respiration sensor for sleep apnea detection in a clinical setting* (Master's thesis), Oslo, Norway: University of Oslo, Department of Informatics.

Berry, R. B., Brooks, R., Gamaldo, C. E., Harding, S. M., Marcus, C., Vaughn, B. V., et al. (2012). The AASM manual for the scoring of sleep and associated events. In *Rules, terminology and technical specifications, darien, illinois, american academy of sleep medicine, Vol. 176* (p. 2012).

Chadha, T., Watson, H., Birch, S., Jenouri, G., Schneider, A., Cohn, M., et al. (1982). Validation of respiratory inductive plethysmography using different calibration procedures. *American Review of Respiratory Disease*, *125*(6), 644–649.

Chang, Y., Xu, L., Han, F., Keenan, B. T., Kneeland-Szanto, E., Zhang, R., et al. (2019). Validation of the nox-T3 portable monitor for diagnosis of obstructive sleep apnea in patients with chronic obstructive pulmonary disease. *Journal of Clinical Sleep Medicine*, *15*(4), 587–596.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Cohn, M., Rao, A., Broudy, M., Birch, S., Watson, H., Atkins, N., et al. (1982). The respiratory inductive plethysmograph: a new non-invasive monitor of respiration.. *Bulletin Europeen de Physiopathologie Respiratoire*, *18*(4), 643.

Collop, N. A., Tracy, S. L., Kapur, V., Mehra, R., Kuhlmann, D., Fleishman, S. A., et al. (2011). Obstructive sleep apnea devices for out-of-center (OOC) testing: technology evaluation. *Journal of Clinical Sleep Medicine*, *7*(5), 531–548.

Dehkordi, P., Marzencki, M., Tavakolian, K., Kaminska, M., & Kaminska, B. (2012). Monitoring torso acceleration for estimating the respiratory flow and efforts for sleep apnea detection. In *2012 annual international conference of the ieee engineering in medicine and biology society* (pp. 6345–6348). IEEE.

ElMoaqet, H., Eid, M., Glos, M., Ryalat, M., & Penzel, T. (2020). Deep recurrent neural networks for automatic detection of sleep apnea from single channel respiration signals. *Sensors*, *20*(18), 5037.

Faust, O., Acharya, U. R., Ng, E., & Fujita, H. (2016). A review of ECG-based diagnosis support systems for obstructive sleep apnea. *Journal of Mechanics in Medicine and Biology*, *16*(01), Article 1640004.

Gutiérrez-Tobal, G. C., Álvarez, D., Crespo, A., del Campo, F., & Hornero, R. (2018). Evaluation of machine-learning approaches to estimate sleep apnea severity from at-home oximetry recordings. *IEEE Journal of Biomedical and Health Informatics*, *23*(2), 882–892.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., et al. (2017). Deep learning scaling is predictable, empirically. arXiv preprint arXiv:1712.00409.

Hrubos-Strøm, H., Randby, A., Namtvedt, S. K., Kristiansen, H. A., Einvik, G., Benth, J., et al. (2011). A norwegian population-based study on the risk and prevalence of obstructive sleep apnea the akershus sleep apnea project (ASAP). *Journal of Sleep Research*, *20*(1pt2), 162–170.

Huang, Q., Qin, Z., Zhang, S., & Chow, C. (2008). Clinical patterns of obstructive sleep apnea and its comorbid conditions: a data mining approach. *Journal of Clinical Sleep Medicine: JCSM: Official Publication of the American Academy of Sleep Medicine*, *4*(6), 543–550.

Kristiansen, S., Andersen, M. H., Goebel, V., Plagemann, T., Traaen, G. M., Øverland, B., et al. (2021). Evaluating a low-cost strain gauge breathing sensor for sleep apnea detection at home. In *2021 IEEE international conference on communications workshops (ICC Workshops)* (pp. 1–6).

Kristiansen, S., Hugaas, M. S., Goebel, V., Plagemann, T., Nikolaidis, K., & Liestøl, K. (2018). Data mining for patient friendly apnea detection. *IEEE Access*, *6*, 74598–74615.

Kristiansen, S., Nikolaidis, K., Plagemann, T., Goebel, V., Traaen, G. M., Øverland, B., et al. (2021). Machine learning for sleep apnea detection with unattended sleep monitoring at home. *ACM Transactions on Computing for Healthcare*, *2*(2), 1–25.

Kristiansen, S., Traaen, G. M., Øverland, B., Plagemann, T., Gullestad, L., Akre, H., et al. (2020). Comparing manual and automatic scoring of sleep monitoring data from portable polygraphy. *Journal of Sleep Research*, Article e13036.

Lin, Y.-Y., Wu, H.-T., Hsu, C.-A., Huang, P.-C., Huang, Y.-H., & Lo, Y.-L. (2016). Sleep apnea detection based on thoracic and abdominal movement signals of wearable piezoelectric bands. *IEEE Journal of Biomedical and Health Informatics*, *21*(6), 1533–1545.

Løberg, F., Goebel, V., & Plagemann, T. (2018). Quantifying the signal quality of low-cost respiratory effort sensors for sleep apnea monitoring. In *Proceedings of the 3rd international workshop on multimedia for personal health and health care* (pp. 3–11).

Magalang, U. J., Chen, N.-H., Cistulli, P. A., Fedson, A. C., Gíslason, T., Hillman, D., et al. (2013). Agreement in the scoring of respiratory events and sleep among international sleep centers. *Sleep*, *36*(4), 591–596.

Malhotra, R. K., Kirsch, D. B., Kristo, D. A., Olson, E. J., Aurora, R. N., Carden, K. A., et al. (2018). Polysomnography for obstructive sleep apnea should include arousal-based scoring: an American academy of sleep medicine position statement. *Journal of Clinical Sleep Medicine*, *14*(7), 1245–1247.

Mello, A. A., D Angelo, G., Santos, R. B., Bensenor, I., Lotufo, P. A., Lorenzi-Filho, G., et al. (2022). Influence of the device used for obstructive sleep apnea diagnosis on body position: a comparison between polysomnography and portable monitor. *Sleep Breath*.

Mendonça, F., Mostafa, S. S., Ravelo-García, A. G., Morgado-Dias, F., & Penzel, T. (2018). Devices for home detection of obstructive sleep apnea: A review. *Sleep Medicine Reviews*, *41*, 149–160.

Mendonca, F., Mostafa, S. S., Ravelo-García, A. G., Morgado-Dias, F., & Penzel, T. (2018). A review of obstructive sleep apnea detection approaches. *IEEE Journal of Biomedical and Health Informatics*, *23*(2), 825–837.

Mostafa, S. S., Mendonça, F., G Ravelo-García, A., & Morgado-Dias, F. (2019). A systematic review of detecting sleep apnea using deep learning. *Sensors*, *19*(22), 4934.

Nepal, K., Biegeleisen, E., & Ning, T. (2002). Apnea detection and respiration rate estimation through parametric modelling. In *Proceedings of the IEEE 28th annual northeast bioengineering conference (IEEE Cat. No. 02CH37342)* (pp. 277–278). IEEE.

Nox T3 sleep monitor, nox medical. (2020). http://noxmedical.com/products/nox-t3-sleep-monitor. Accessed: 2020-06-16.

Pombo, N., Garcia, N., & Bousson, K. (2017). Classification techniques on computerized systems to predict and/or to detect apnea: A systematic review. *Computer Methods and Programs in Biomedicine*, *140*, 265–274.

Punjabi, N. M. (2008). The epidemiology of adult obstructive sleep apnea. *Proceedings of the American Thoracic Society*, *5*(2), 136–143.

Sweetzpot. (2020). https://www.sweetzpot.com/. Accessed: September 2020.

Traaen, G. M., Øverland, B., Aakerøy, L., Hunt, T., Bendz, C., Sande, L., et al. (2020). Prevalence, risk factors, and type of sleep apnea in patients with paroxysmal atrial fibrillation. *IJC Heart & Vasculature*, *26*, Article 100447.

Tsouti, V., Kanaris, A., Tsoutis, K., & Chatzandroulis, S. (2020). Development of an automated system for obstructive sleep apnea treatment based on machine learning and breath effort monitoring. *Microelectronic Engineering*, Article 111376.

Tumer, K., & Ghosh, J. (1996). Estimating the Bayes error rate through classifier combining. *2*, In *Proceedings of 13th international conference on pattern recognition* (pp. 695–699). IEEE.

Uddin, M., Chow, C., & Su, S. (2018). Classification methods to detect sleep apnea in adults based on respiratory and oximetry signals: a systematic review. *Physiological Measurement*, *39*(3), 03TR01.

Van Steenkiste, T., Groenendaal, W., Dreesen, P., Lee, S., Klerkx, S., De Francisco, R., et al. (2020). Portable detection of apnea and hypopnea events using bio-impedance of the chest and deep learning. *IEEE Journal of Biomedical and Health Informatics*.

Wang, Y., Huang, M., Zhu, X., & Zhao, L. (2016). Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 606–615).

Weiss, G. M., & Provost, F. (2001). The effect of class distribution on classifier learning: an empirical study.

Young, T., Skatrud, J., & Peppard, P. E. (2004). Risk factors for obstructive sleep apnea in adults. *Jama*, *291*(16), 2013–2016.