



Obstructive Sleep Apnea Prediction: A Comprehensive Review and Comparative Study

Huynh Thi Khanh Chi¹ · Amonae Dabbs-Brown² · Anna Jurek-Loughrey¹ · James Mulhall¹ · Tuan Dung Pham¹ · Ngoc Phu Doan¹ · Viet Hung Tran¹ · Zichi Zhang¹ · Xuan Hoang Nguyen⁵ · Yimeng An^{1,3} · Peixin Li^{1,3} · Phi Hung Nguyen¹ · Thi Linh Hoang⁶ · Xinming Shi¹ · Hans Vandierendonck¹ · Sebastien Bailly⁴ · Jean-Louis Pépin⁴ · Thai Son Mai¹

Received: 11 April 2025 / Revised: 30 September 2025 / Accepted: 25 November 2025 /
Published online: 23 January 2026
© The Author(s) 2025

Abstract

Obstructive Sleep Apnea (OSA) is a highly prevalent sleep disorder linked to considerable public health burdens and comorbidities. However, its heterogeneous presentation and the limited accessibility of traditional diagnostic tools such as polysomnography (PSG) lead to widespread underdiagnosis. As a result, artificial intelligence (AI) approaches, including machine learning (ML) and deep learning (DL) models, have attracted attention as an alternative pathway to detection. This paper first provides a comprehensive review of AI-driven OSA diagnosis, covering different diagnosis problems, input-data types, data biases, pre-processing techniques, and model performance. We then leverage the largest clinical dataset used in OSA prediction to date, approximately 110,000 patients with 22,000 having complete entries for all 50 features, to systematically compare the performance of 39 ML/DL models. Our findings highlight the challenging nature of OSA prediction, with accuracies ranging from 29.66% to 46.9% for 4-class prediction and 46.04% to 87.18% for binary tasks. DL models such as DANet and GATE scored highest, whereas ensemble approaches such as LGBM and AdaBoost displayed more consistent performance across folds. However, as severe cases of OSA are easier to predict and over-represented in datasets, accuracy alone is insufficient for model evaluation and we explore a variety of metrics. Finally, imbalance correction and feature selection improved weaker models, but had only marginal effects on the best-performing models. Looking forwards, the development of more sophisticated and tailored DL models and large, high-quality datasets may help to break current performance barriers. We hope that our work can attract more attention to this challenging but interesting research problem.

Huynh Thi Khanh Chi, Amonae Dabbs-Brown, Anna Jurek-Loughrey, James Mulhall and Thai Son Mai have equally contributed to this work.

Editors: Riccardo Guidotti, Anna Monreale, Dino Pedreschi.

Extended author information available on the last page of the article

Keywords Obstructive sleep apnea (OSA) · Machine learning · Deep learning · Supervised learning

1 Introduction

Sleep apnea is one of the most common sleep disorders in the world, affecting approximately one billion people (Benjafield et al., 2019). It occurs when a person repeatedly stops and starts breathing during sleep (National Heart Lung & Blood Institute, 2022), and is divided into three main forms (Mayo Clinic, 2020; Wang et al., 2013). The first type, central sleep apnea, occurs when the brain stops sending signals to the muscles that control breathing. The second type, obstructive sleep apnea (OSA), is caused by a collapse or blockage of the upper airway during sleep. The last type, complex sleep apnea, is a distinct form of central sleep apnea that emerges or persists in OSA patients that are treated with a continuous positive airway pressure (CPAP) device. Among them, OSA is the most prevalent and is the main focus of this study.

1.1 The Health and Economic Burdens of OSA

OSA has serious health consequences, including hypoxia, hypercapnia, heart rate and blood pressure fluctuations, increased sympathetic activity, cortical arousal, or arousal from sleep (Young et al., 2004). It can lead to fragmented, non-restorative sleep, causing daytime sleepiness, fatigue, poor concentration, and irritability (Engleman & Douglas, 2004; Knauert et al., 2015). Untreated OSA is a major public health concern, linked to a higher risk of conditions such as hypertension, diabetes, stroke, heart diseases, chronic obstructive pulmonary disease (COPD), kidney disease, and increased risk of mortality (Peppard et al., 2000; Knauert et al., 2015; Punjabi et al., 2009; Gottlieb & Punjabi, 2020; Bonsignore et al., 2019). The condition also impairs cognitive function (Engleman & Douglas, 2004) and is associated with an increased risk of dementia (Kim et al., 2013). Moreover, the depression rate of people with OSA is much higher than those without OSA, especially in women (Chen et al., 2013; Knauert et al., 2015).

The global impact of OSA is significant and growing, affecting an estimated 14% of men and 5% of women worldwide (Kapur et al., 2017; Gottlieb & Punjabi, 2020; Randerath et al., 2018; Tan et al., 2013). However, despite its severity, more than 80% of patients remain undiagnosed (Bailly et al., 2016). This lack of diagnosis poses a substantial economic and social burden (Knauert et al., 2015). For example, drivers with OSA are up to 6.3 times more likely to be in a traffic accident than healthy drivers (Teran-Santos et al., 1999; Knauert et al., 2015). In 2000, around 800,000 traffic accidents in the US, costing \$15.9 billion for related consequences and 1,400 lives, were related to drivers with OSA. Treating all of these drivers may consume \$3.18 billion, but it would be helpful to save \$11.1 billion in accident costs and 980 lives each year. Hence, it is crucial for effective and early diagnosis and treatment of OSA (Knauert et al., 2015).

1.2 Challenges in OSA Diagnosis

The standard method for diagnosing OSA is polysomnography (PSG) which measures the Apnea-Hypopnoea Index (AHI) in a sleep laboratory (Young et al., 2004). An AHI of 5 or more indicates OSA, with higher scores signifying greater severity (Young et al., 2004; Gottlieb & Punjabi, 2020). However, PSG is highly invasive, expensive, and limited by the small number of available sleep laboratories and specialists (Punjabi, 2008; Young et al., 2004; Kapur et al., 2017). This makes it inaccessible for many, leaving a significant gap in diagnosis. However, simpler methods, such as sleep history questionnaires, are often inaccurate (Kapur et al., 2017; Gottlieb & Punjabi, 2020). Hence, there is a need for faster, cheaper, and more efficient diagnosis tools for OSA (Randerath et al., 2018).

1.3 Machine Learning for OSA Assessment

Artificial Intelligence (AI), including Machine Learning (ML) and Deep Learning (DL), is being increasingly applied to OSA severity assessment in recent research efforts, e.g. Bailly et al. (2016); Mai et al. (2018); Lacedonia et al. (2016); Wu et al. (2016); Varon et al. (2015); Huang et al. (2020a). These works employ very diverse approaches from conventional ML models (e.g., Random Forest, Gradient Boosting, XGBoost (Chen & Guestrin, 2016), Support Vector Machine (SVM), and Hist Gradient Boosting) to DL models (e.g., LSTM (Van Steenkiste et al., 2018), 1D-CNN (Sharma et al., 2022b), and ResNet (Shanmugham et al., 2021)). These models have been applied to analyze various types of data, such as electrocardiography (ECG) (Hassan & Haque, 2017), oxygen saturation (SpO₂) (Sharma et al., 2022a), human facial scans (Islam et al., 2018), tracheal respiratory movements (Hafezi et al., 2020), chin electromyography signals (EMG) (Moradhasel et al., 2023), or patient clinical data from Electronic Health Records (EHR).

Among the mentioned data types, EHR data is particularly attractive for its ease of collection, while other data require tools and machinery that may not be convenient and accessible for patients (Huang et al., 2020a; Wu et al., 2016; Mencar et al., 2020; Rodrigues Jr et al., 2020). However, it presents unique challenges for ML models. EHR data is often non-standardised across healthcare institutions, with numerous missing data points and a bias toward patients with more severe conditions. This necessitates extensive data pre-processing, including missing data imputation, imbalance correction, and feature selection (Rodrigues Jr et al., 2020; Wu et al., 2016). The sizes of the datasets also vary significantly between studies, ranging from a few to thousands of subjects, e.g., 4 in Javaid et al. (2015) and 6,875 in Huang et al. (2020a). The smaller the number of samples, the more likely the models are to be biased, thus reducing their applicability in reality. Additionally, reported prediction accuracies vary considerably. For example, Li et al. (2018) report 100% accuracy with the single-lead ECG signal for detecting OSA ($AHI \geq 5$) in 313 patients, while Ramesh et al. (2021) report 68.06% for the same problem but with 1,479 patients and common clinical features obtained from EHR. This inconsistency highlights the complexity and ongoing challenges in using ML for reliable OSA assessment.

1.4 Our Contributions

Due to the rapidly growing prevalence of OSA, its severe effects, and its huge potential damage to society and the economy, it is crucial to have early and effective diagnoses for a wide range of OSA patients and to better understand its complexity. This paper aims at two main targets presented in two key parts.

In **Part I: Literature Review**, we provide a comprehensive review on different DL and ML approaches for OSA severity assessments. Though there exist other surveys in the literature such as Xu et al. (2022); Mostafa et al. (2019), they all focus on one special aspect. For example, Xu et al. (2022) cover all works related to OSA prediction via PSG signals such as ECG or SpO₂, while Mostafa et al. (2019) aim to review all DL-related techniques. Different from these works, our review takes a broader scope, covering different diagnosis problems (including 4 severity class predictions, binary severity predictions and direct AHI prediction), types of data (including clinical data, PSG signals and non-PSG data) and their relevant features, different learning techniques (ML and DL), and their prediction outcomes. Hence, it provides a wider overview of the field and is intended to be useful to other researchers for their future studies.

In **Part II: Comparative Experiment**, though OSA severity assessments via EHR clinical data are a very attractive approach as mentioned above with many published works such as Wu et al. (2016); Mencar et al. (2020); Rodrigues Jr et al. (2020), there remain many open questions and limitations such as small and biased training/evaluation data, limited clinical features, inconsistent reported results, and lack of DL methods (c.f. Sect. 3.9.1 for details). Hence, in this paper, we do not focus on providing new methods, but we conduct a comprehensive comparative study to provide deep insights into all aspects of OSA severity assessments using EHR clinical data. Our studies are built upon a large and unique dataset with 110,000 patients and hundreds of EHR clinical features collected from the French national registry of sleep apnea (OSFP), which can reduce data biases. Moreover, we study a wide range of ML and DL methods, including those commonly used and those that have not been studied in OSA research before. Specifically, we evaluate 39 ML and DL models, 17 feature sets (selected either automatically by different algorithms or manually from prior studies), 4 imbalance correction techniques, 16 data imputation methods, and 3 prediction tasks (binary severity classification, 4-level severity classification, and direct AHI prediction). This results in 127,296 unique experimental configurations. Therefore, our study portrays a broad and detailed view of the performance of OSA predictions from EHR data as shown in Sect. 4.

Our studies show that effectively predicting OSA from EHR data is a very challenging problem with particularly low prediction accuracy and biased outcomes towards higher severity levels (c.f. Sect. 4.3), while common data preprocessing techniques like imputation (c.f. Sect. 4.6), imbalance correction (c.f. Sect. 4.5), and feature selection (c.f. Sect. 4.4) do not help to improve performance in general. Increasing training data can make it even worse due to many missing data points (c.f. Sect. 4.6). However, the slightly better performance of DL methods compared to conventional ML techniques raises hopes for further performance improvements subject to larger and cleaner data and more sophisticated DL models and learning strategies specifically designed for OSA severity prediction problems (c.f. Sect. 4.8.1).

We hope that our work can help attract more attention from ML/AI communities to this important, interesting but particularly challenging problem to break the current limitations and performance barriers.

Outline. The remainder of the paper is structured as follows. The general background of OSA disease can be found in Sect. 2, including its risk factors, symptoms, comorbidities, and current diagnostic processes. The main contributions of the paper are divided into two parts: (i) In Part I (Sect. 3), we survey potential ML and DL approaches for OSA diagnosis from both previous works and the perspective presented in this paper and (ii) Extensive experiments are conducted using a wide range of modern ML and DL models on our dataset and the results analysis can be found in Part II (Sect. 4). Finally, conclusions are drawn in Sect. 5.

2 Background

Figure 1 illustrates the overall background of OSA, including risk factors, symptoms, comorbidities, and diagnosis. Table 6 also shows 49 common OSA-related factors used in previous works (Mencar et al., 2020; Wu et al., 2016; Huang et al., 2020a) such as age, height, etc., grouped into 5 categories including: demographics, measurements, questionnaires, comorbidities, and symptoms.

2.1 Risk Factors

Several key factors are associated with obstructive sleep apnea (OSA), often alongside its primary indicator, snoring. These include gender, age, weight, lifestyle, and family history (Gottlieb & Punjabi, 2020). Firstly, men are two to three times more likely to snore and develop OSA (Semelka et al., 2016). This may be attributed to sex-related differences in airway pathophysiology, as studies have shown a positive correlation between neck circumference and increased AHI (Appleton et al., 2018; Ahbab et al., 2013). Secondly, OSA, like snoring, becomes more common with age, especially over 60 (Franklin & Lindberg, 2015). Thirdly, obesity can contribute to OSA by mass loading the upper airway, and it has been estimated that over half of moderate to severe cases of OSA are due to a body-mass index (BMI) ≥ 25 kg/m² (Young et al., 2005). Fourthly, both cigarette smoking and alcohol use are also linked to OSA (Gottlieb & Punjabi, 2020; Young et al., 2005). Cigarette smoking may cause airway inflammation, while alcohol intake may reduce motor output to the upper airway, though these attributes are also linked to others, such as obesity (Young et al., 2005). Finally, genetic factors that may contribute to OSA include those that predispose people to

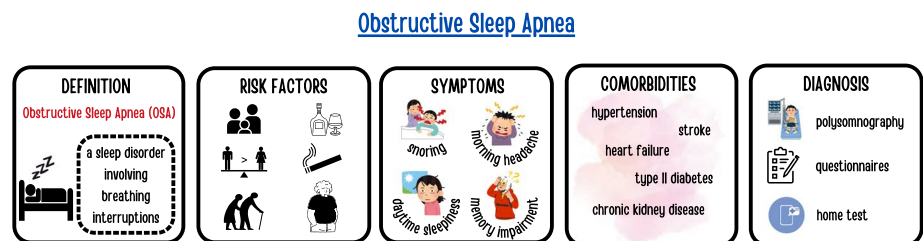


Fig. 1 Overall background of the OSA disease

obesity or inflammation and those that affect the face and skull structure (National Heart Lung & Blood Institute, 2022).

2.2 Symptoms

Traditional signs and symptoms of OSA include snoring, choking or gasping during sleep, nocturia, morning headache, and excessive daytime sleepiness (National Heart Lung & Blood Institute, 2022). These sleep-related symptoms may also lead to depression, anxiety, and decreased attention, concentration, motor skills, and memory (Engleman & Douglas, 2004; Chen et al., 2013; Knauert et al., 2015). However, recent studies have shown that up to 15–50% of patients diagnosed with OSA have self-reported excessive daytime sleepiness, a cardinal symptom of OSA (Gottlieb & Punjabi, 2020; Arnardottir et al., 2015). In addition, although habitual snoring is present in 50–60% of those with OSA, a systematic review concluded that it is a less reliable indicator of OSA than nocturnal gasping or choking, as it is not specific to the disorder (Gottlieb & Punjabi, 2020). Importantly, therefore, OSA patients are heterogeneous in terms of symptoms, comorbidities, physiological traits linked to pathogenesis, and the expression of the disorder (Zinchuk & Yaggi, 2019).

2.3 Comorbidities

OSA is a common comorbidity for several chronic conditions, including cardiovascular disease, metabolic disorders, and respiratory illnesses (COPD) (Bonsignore et al., 2019; Chiang et al., 2017). It is also highly prevalent in patients with resistant hypertension, atrial fibrillation, and type 2 diabetes, with some studies estimating its presence in over 70% of these populations (Gottlieb & Punjabi, 2020; Bonsignore et al., 2019; Chiang et al., 2017; Javaheri et al., 2017). The risk of developing hypertension rises with OSA severity, which was reported by a systematic review for an odds ratio of 2.8 for the association with resistant hypertension and a graded increase in risk with essential hypertension (mild OSA: 1.2, moderate: 1.3, severe: 1.6) (Hou et al., 2018). The link is also strong with coronary artery disease and stroke (Franklin & Lindberg, 2015).

Recently, there has been an increased interest in studying the role of OSA as an independent, modifiable risk factor, as the number of comorbidities is higher in patients with OSA relative to other sleeping disorders. One possible reason for this is that the frequent drops in blood oxygen saturation during apnea-hypopnea episodes trigger an autonomic nervous system response, activating the sympathetic nervous system. This physiological chain reaction could then set off a cascade of events culminating in disease development. Given this, treating OSA is increasingly seen as a way to reduce health complications and has been associated with fewer cardiovascular events in patients with multiple comorbidities (Bonsignore et al., 2019).

2.4 Diagnosis

Diagnosing OSA begins with a comprehensive sleep evaluation, as recommended by the American Academy of Sleep Medicine (AASM). This includes a detailed sleep history and a physical examination of the patient's respiratory, cardiovascular, and neurological systems (Kapur et al., 2017). The gold standard for diagnosing OSA is polysomnography (PSG).

This is a sleep study performed in a laboratory over 8 h and requires that a minimum of 22 wires be attached to a patient's body (Penzel et al., 2016; Jeon et al., 2020). Twelve different signals are monitored during PSG, such as electrocardiography (ECG), electromyography (EMG), electroencephalography (EEG), electrooculography (EOG), and oxygen saturation (SpO₂) (Penzel et al., 2016). Abdominal and thoracic movement, body posture, snoring level, and REM (rapid eye-movement) together with non-REM sleep are also monitored. During a PSG, a patient's AHI is recorded. The severity levels of OSA diagnosis are commonly determined via AHI per hour as *None* ($AHI < 5$), *Mild* ($15 > AHI \geq 5$), *Moderate* ($30 > AHI \geq 15$), and *Severe* ($AHI \geq 30$). The results of a PSG determine the need for supplemental oxygen and/or positive airway pressure titration.

Alternatively, home sleep apnea tests (HSATs) can be used to confirm a diagnosis of OSA. Though HSATs are more easily accessible and less costly, they confer an increased risk of technical failure and user error (Kapur et al., 2017). There are several tools to aid in the diagnosis of OSA including sleep history questionnaires such as the STOP-BANG Questionnaire (Chung et al., 2016), Epworth Sleepiness Scale (Johns, 1991), and Pichot's Fatigue Scale (Pichot & Brun, 1984), and clinical tools such as upper airway examinations (Rundo, 2019; Stanford Health Care (SHC) - Stanford Medical Center, 2017). The AASM, however, discourages the use of clinical tools, questionnaires, and prediction algorithms to diagnose OSA in the absence of PSG or HSATs (Kapur et al., 2017).

Prediction algorithms using ML or DL can be less invasive and lower the cost and burdens to patients and clinicians that HSATs and PSG generate, leading to earlier and easier diagnosis as we will study in Sect. 3 below.

3 Part I: Literature Review

In this Literature Review Part, we provide a comprehensive literature review on OSA severity study. Figure 2 illustrates common frameworks used in the literature for OSA severity assessments.

3.1 Prediction Problems

Most existing works either predict AHI values directly (Rodrigues Jr et al., 2020; Mencar et al., 2020) or predict severity levels (i.e. OSA diagnosis). These severity level prediction problems can be in the form of all classes prediction problems (Rodrigues Jr et al., 2020;

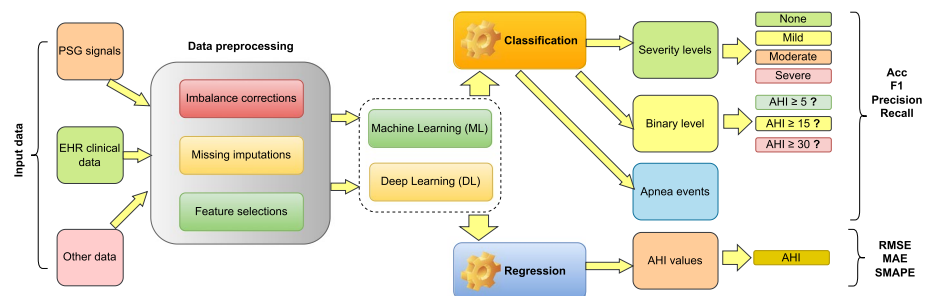


Fig. 2 Overall framework for OSA severity predictions in the literature

Mencar et al., 2020), or binary classes prediction problems (Huang et al., 2020a; Wu et al., 2016; Keshavarz et al., 2020; Ustun et al., 2016), using AHI as level thresholds with common values {5, 15, 30}, as indicated in Sect. 2. For example, Wu et al. (2016) report a prediction accuracy of 75.60% when classifying $AHI \geq 15$ or not with a Neural Fuzzy Evaluation System (NFES) and a RMSE of 16.61 when predicting AHI value directly on a dataset with 150 patients. The remaining studies focus on detecting apnea events from PSG signals, such as the work conducted by Mahmud et al. (2021).

3.2 Input Data

For OSA diagnosis, input data can be diverse, with examples including PSG-related signals (Varon et al., 2015), EHR clinical data (Wu et al., 2016; Mencar et al., 2020), and other data such as sensor data collected from wearable devices, cameras, and microphones (Ramachandran & Karupiah, 2021) or images (Qiu et al., 2026).

3.2.1 PSG Signals

These include all signals related to PSG such as ECG, EMG, EEG, and SpO₂ data, or their derivatives including heart rate variability (HRV), ECG-derived respiration (EDR) (Jayawardhana & de Chazal, 2017; Burgos et al., 2009; Penzel et al., 2000; Goldberger et al., 2000; Gutiérrez-Tobal et al., 2018; Varon et al., 2015; Hassan & Haque, 2017). Published datasets that were acquired from the PSG method can be found in Table 1, while the involvement of ML in related works with these datasets can be found in Table 2. Models using PSG signals as features may use classifiers to either detect sleep apnea events or diagnose sleep apnea severity. Of the PSG data, SpO₂ and ECG signals constitute the bulk of features used due to their correlation with apneic events (Jayawardhana & de Chazal, 2017).

For SpO₂, Jayawardhana and de Chazal (2017) were able to detect sleep apnea events using the photoplethysmography (PPG) measurements (a SpO₂ derivative) of 52 patients with 87% accuracy. Their model is built using a Linear Discriminant Analysis classifier which separates classes based on a linear combination of features. Burgos et al. (2009) describe a method to track apnea events in real-time using SpO₂ signals available in the Apnea-ECG Database (Penzel et al., 2000; Goldberger et al., 2000) to train an ensemble tree-based classifier. This model achieves an accuracy of 93%. Gutiérrez-Tobal et al. (2018) also use SpO₂ measurements to classify sleep apnea in patients. In their study, they use single-channel recordings from 320 patients to classify sleep apnea syndrome severity with 70% accuracy using an AdaBoost model built with linear discriminants.

Efforts to detect sleep apnea based on ECG signals primarily focus on data derived from a single-lead ECG signal. For example, Varon et al. (2015) present a methodology to detect sleep apnea on a minute-by-minute basis using a least-squares SVM classifier and features developed within the study. The classifier achieves an accuracy of 85% when predicting AHI and 100% when distinguishing between AHI greater or less than 10. Hassan and Haque (2017) create a novel ML algorithm, RUSBoost, to classify patients as non, borderline, and apnea with over 85% accuracy. The algorithm combines random undersampling with boosting to adjust for the majority of samples coming from the apnea class. Zarei and Asl (2020) use an RF model to distinguish between patients with and without OSA with 97% accuracy using a spectral autocorrelation function to extract features from a single-lead ECG signal.

Table 1 Public PSG datasets for OSA detection

Dataset	# Reco.	# Anno.	% OSA	# Subj.	AHI	BMI (kg/m ²)	Age (years)	Gen- der (% female)	Papers using this dataset
AED (Penzel et al., 2000; Goldberger et al., 2000)	70	35	66%	32	(0, 83)	(19.2, 45.33)	(27, 63)	22%	(Li et al., 2018; Bernardini et al., 2021; Khandoker et al., 2009; Hemrajani et al., 2025; Hu et al., 2025; Zhang et al., 2025b; Padovano et al., 2025)
SVUH- UCD (Heneghan, 2011; Gold- berger et al., 2000)	25	—	—	25	(1.7, 90.9)	(25.1, 42.5)	(28, 68)	16%	(Hassan & Haque, 2017; Xie & Minn, 2012; Mahmud et al., 2021; Zhang et al., 2025b; Padovano et al., 2025)
PNC-MGH (Ghassemi et al., 2018)	1,983	994	81%	1,983	19 (14.4)	33 (7.6)	55 (14.4)	35%	(Liu et al., 2020; Bark et al., 2023; Mahmud et al., 2021; Yousefi et al., 2025; Park et al., 2025)
SHHS-1 (Zhang et al., 2018; Quan et al., 1997)	5,793	—	—	5,793	(5, 30)	(25, 31)	(55, 74)	47%	(Deviaene et al., 2018; Sharma et al., 2022b; Liang, 2023; Alarcón et al., 2024; Hu et al., 2025; Park et al., 2025)

Table 1 (continued)

Dataset	# Reco.	# Anno.	% OSA	# Subj.	AHI	BMI (kg/m ²)	Age (years)	Gen- der (% female)	Papers using this dataset
SHHS-2 (Zhang et al., 2018; Quan et al., 1997)	2,651	–	–	2,651	(7, 25)	(25, 31)	(60, 76)	46%	(Sharma et al., 2022a; Deviaene et al., 2018; Alarcón et al., 2024; Hu et al., 2025)
MESA (Zhang et al., 2018; Chen et al., 2015)	2,273	2,056	861	4,077	14.5 (15.7)	28.5 (6.9)	67 (-)	48%	(Olsen et al., 2020; Ben Sason et al., 2023; Alarcón et al., 2024; Agrawal et al., 2025; Hu et al., 2025)

#Reco.: Number of Records, #Anno.: Number of Annotated Records, Subj.: Subjects

Using beat-to-beat (RR) interval-based features and SVM, Yılmaz et al. (2010) detect sleep apnea events with 89% accuracy.

Though other signals like EMG, HR, Airflow or Respiratory are less frequently used like SpO₂ and ECG, they acquire comparable OSA prediction accuracy. For example, Moradhasel et al. (2023) report an accuracy of 99.1% when predicting apnea events using EMG and EfficientNet architecture. Or Mahmud et al. (2021) acquire an accuracy of 93.2% for the same event detection task with EEG and BiLSTM architecture.

From the literature, it can be seen that models using PSG signals perform relatively well at OSA diagnosis and classification. However, this comes at the expense of requiring inconvenient and often inaccessible diagnostic facilities. Moreover, recent approaches usually do not rely on a single PSG signal but are aimed at combining them as multi-modal data to complement each other, thus enhancing prediction performance. For example, Hemrajani et al. (2023) combine ECG, Airflow and Abdominant respiratory signals in the 1D-CNN+Bi-GRU model to have 95.1% OSA event detection accuracy compared with 91.9% if using only ECG in Hassan and Haque (2017).

3.2.2 Clinical Data

Other attempts have also been made to classify OSA using EHR clinical data as shown in Table 4. For example, Ustun et al. (2016) uses an improved Supersparse Linear Integer Model (SLIM) to develop an OSA diagnostic screening tool based on readily available clinical information including age, sex, BMI, and hypertension status. The tool has a sensitivity of 64% and a specificity of 77%. Mencar et al. (2020) attempted to both predict AHI and classify OSA severity using 3–8 features gathered before PSG, including BMI, age, sex, and Epworth Sleepiness Scale results. The study tests seven types of classification models to predict severity and five types of regression models to predict AHI, and noted that RF is optimal for classification while SVM is best for regression. The highest average accuracy achieved is 44.7%. Wu et al. (2016) use a similar feature set with the addition of blood pres-

Table 2 Supervised learning in OSA detection using PSG signals

Dataset	PSG Signal(s)	Classifier (Best)	Acc.	Sens.	Spec.	S/A
AED	ECG	DNN-ANN/SVM-HMM + Decision Fusion (Li et al., 2018)	100%	100%	100%	A
		CNN-LSTM (Bernardini et al., 2021)	100%	100%	100%	A
		LS-SVM (Sharma & Sharma, 2016)	97.1%	95.8%	100%	A
		HMM-SVM (Song et al., 2015)	97.1%	95.8%	100%	A
		RF (Nishad et al., 2024)	92.8%	91.0%	93.9%	A
		SVM (Viswabhargav et al., 2019)	—	85.43%	92.6%	A
		kNN (Sunil Kumar & Kanhangad, 2017)	89.8%	88.5%	90.6%	A
		AdaBoost (Hassan, 2016)	87.3%	82.0%	90.7%	A
		SVM (Ivanko et al., 2020)	98.7%	—	—	A
		RF (Zarei & Asl, 2020)	97.1%	95.7%	100%	A
		LR (Alvarez et al., 2010)	89.7%	92.0%	85.4%	A
		RF (Deviaene et al., 2018)	91.3%	98.9%	86.6%	A
		PNN (Morillo & Gross, 2013)	93.8%	92.4%	95.9%	A
		Bagging-ADTree (Burgos et al., 2009)	93.0%	92.4%	93.5%	A
PNC-MGH	EEG	BiLSTM (Mahmud et al., 2021)	93.2%	91.7%	93.7%	A
		XGBoost (Liu et al., 2020)	91.3%	95.1%	88.5%	A
	ECG + Respiratory	ResNet(Shanmugham et al., 2021)	77.0%	—	—	A
	ECG + SaO2	1D-CNN-LSTM +Selective Prediction (Bark et al., 2023)	90.3%	91.3%	89.2%	A
	ECG	LS-SVM (Varon et al., 2013)	85.1%	70.2%	91.1%	A
	ECG	RUBoost (Hassan & Haque, 2017)	91.9%	90.4%	92.7%	A
	ECG + SpO2	Bagging-REPTree(Xie & Minn, 2012)	84.4%	79.8%	85.9%	A
	EEG, EOG, EMG, Air-flow, SpO2	CNN+BiLSTM (Yousefi et al., 2025)	81.3%	71.2%	81.9%	A
	ECG, Air-flow, AbdRes	1D-CNN+Bi-GRU (Hemrajani et al., 2025)	95.1%	96.1%	93.4%	A
SHHS-1	Respiratory	LSTM (Van Steenkiste et al., 2018)	77.2%	62.3%	80.3%	A
	SpO2	1D-CNN (Sharma et al., 2022b)	84.3%	68.9%	86.4%	A
		SVM (Liang, 2023)	82.6%	82.0%	82.7%	A
AED + SVUH-UCD	ECG	SVM (Chen et al., 2014)	97.4%	99.0%	92.9%	A
	ECG, SpO2	Transformer (Zhang et al., 2025b)	95.0%	—	—	A
AED + SVUH-UCD + In-house	ECG	SVM (Khandoker et al., 2008)	92.9%	—	—	S
SHHS-1 + SHHS-2	SpO2, Respiratory	RUSBoost (Sharma et al., 2022a)	88.3%	—	—	A
SHHS-1 + PNC-MGH	EEG	DistillSleep (Transformer) (Park et al., 2025)	—	—	—	A
SSSH-1 + MESA	SpO2, HR, AbdRes	UNet (Alarcón et al., 2024)	88.0%	88.0%	89.0%	A
	Flow, Chest, SpO2	EAT-Net (Hu et al., 2025)	87.5%	84.4%	92.2%	A
MESA	PPG	MLSTM-FCN (Agrawal et al., 2025)	93.4%	94.0%	91.0%	A

Table 2 (continued)

Dataset	PSG Signal(s)	Classifier (Best)	Acc.	Sens.	Spec.	S/A
Others	EMG	EfficientNet (Moradhasel et al., 2023)	99.1%	99.2%	98.5%	A
	EEG, EMG, Sleep stage labels	RF (Zhang et al., 2025a)	71.1%	52.1%	82.0%	A

Acc.: Accuracy, Sens.: Sensitivity, Spec.: Specificity, S/A: if it is Severity prediction (S) or Apnea Event detection (A)

sure (BP) measurements to develop a neural fuzzy evaluation system (NFES) capable of distinguishing between patients with $AHI < 15$ and $AHI \geq 15$ with 76% accuracy. Huang et al. (2020a) take a different approach, applying an SVM model in a stepwise model to distinguish between patients with $AHI \geq 5/h$, $\geq 15/h$, and $\geq 30/h$. The model is applied to the entire dataset and four subgroups categorised by gender and age (< 65 versus ≥ 65 years old). The sensitivities for groups in the complete dataset with increasing AHI are 74.14%, 75.18%, and 70.26%, respectively while the specificities were 74.71%, 68.73%, and 70.30%, respectively. Interestingly, within the subgroups, the model performs best for males < 65 years old and modestly for females ≥ 65 years old. Rodrigues Jr et al. (2020) perform an extensive study testing several feature sets including demographic, symptom, and clinical measurement data on over 60 classification and evaluation algorithms. The study demonstrates that features related to age, BMI, and sleep habits were most associated with OSA severity. The top-performing model during the study is Extra Trees which results in a sensitivity of 64% and a specificity of 83%.

In general, when it comes to clinical data, the studies shown here highlight that performance generally worsens compared to the PSG-based models.

3.2.3 Other Data

Although physiological data, or sleep signals, are usually measured during a PSG, they can also be obtained using biomedical and environmental sensors or vision-based systems such as MRI images (Shahid et al., 2015), cameras, wearable devices, microphones, and inertial motion units (Ramachandran & Karuppiah, 2021) as shown in Table 3. Sensor- and vision-based models sometimes reach accuracies much closer to the PSG studies than clinical-based models, providing another viable avenue for further research, especially in multi-modal settings such as Lee et al. (2024); Qiu et al. (2026); Camcı et al. (2017).

3.3 Learning Models

3.3.1 Machine Learning

Conventional ML techniques such as Random Forest (RF) or Linear Regression (LR) are the most widely used models for OSA severity predictions with all kinds of data (Wu et al., 2016; Huang et al., 2020a; Mencar et al., 2020; Ivanko et al., 2020; Alvarez et al., 2010). For example, Nishad et al. (2024) employ Random Forest on ECG signals to predict OSA severity with 92.8% accuracy. Naive Bayes is used to predict binary OSA levels from clinical data in a study by Keshavarz and colleagues (Keshavarz et al., 2020). Among these

Table 3 Supervised learning in OSA detection using Non-PSG signals

Non-PSG signal(s)	#Subj.	Sensor(s)	Classifier (best)	Acc.	S/A
Device's	IR-UWB radar signal(Javaid et al., 2015)	4	IR-UWB radar	LDA	73.1% A
	PSM signals (Azimi et al., 2020)	9	PSM	BiLSTM	95.1% A
	IR-UWB radar signal (Kwon et al., 2021)	36	IR-UWB radar	CNN- LSTM	98.0% S
Sound	Sleep sound signal (Kim et al., 2018)	120	Microphone	LR	92.5% S
	Speech signals (Botelho et al., 2019)	20	Speech recorder	Ensemble- SVM- LDA- kNN	84.4% A
Respiratory	Respiratory signal (Zhuang et al., 2022)	10	FMCW radar	RF	95.5% A
	Respiratory signals (Petrenko, 2020)	8	Accelerometer, pressure transducer	CNN	88.0% A
	Airflow signals (Lakhan et al., 2018)	520	Thermistor	DNN	92.7% S
Image	Depth map of human facial scans (Islam et al., 2018)	69	3D scanner	CNN	67.4% A
Movement	Tracheal respiratory movements (Hafezi et al., 2020)	69	Accelerometer	CNN- LSTM	84.0% S
Combined	Actigraphy, body position, audio, PPG (Behar et al., 2014)	856	Smart phone's application	SVM	92.3% A
	Chest movements, noise level, activity logger (Camci et al., 2017)	4	Phone's accelerometer, microphone	CART	98.2% A
	Breathing sounds, body motions, oximetry (Montazeri Ghahjaverestan et al., 2023)	233	Oximeter, sound, accelerometer	DNN	87.4% A
	Clinical features, portable polysomnography, X-ray and MRI data (Lee et al., 2024)	55	Oximeter, accelerometer	VGG16+LR	91.4% S
	Physiological measures, facial image data, snoring audio (Qiu et al., 2026)	807	Oximeter, sound, accelerometer	MIMAR-OSA	92.2% S

#Subj.: Number of Subjects, Acc.: Accuracy, S/A: if it is Severity prediction (S) or Apnea Event detection (A)

methods, Support Vector Machines (SVM) are the most commonly used and among the top-performing models in many previous works as well (Varon et al., 2015; Yilmaz et al., 2010; Song et al., 2015; Chen et al., 2014; Li et al., 2018; Ivanko et al., 2020; Behar et al., 2014; Huang et al., 2020a; Ramesh et al., 2021; Kim et al., 2021).

3.3.2 Deep Learning

DL models have been successfully applied to the classification of OSA severity in many studies with PSG signals and other data (Nikkonen et al., 2019; Van Steenkiste et al., 2018; Urtnasan et al., 2018; Jiang et al., 2018; Haidar et al., 2017). For example, Nikkonen et al.

(2019) use a multi-layer perceptron (MLP) model with 3 feed-forward layers and backpropagation to classify OSA severity based solely on a patient's blood oxygen saturation signal with 91% accuracy. Van Steenkiste et al. (2018) try to classify OSA severity based on raw physiological respiratory signals using a Long Short-Term Memory (LSTM) model with an average accuracy of 76%. Urtnasan et al. (2018) detect OSA events from single-lead ECG data using a six-layer convolutional neural network (CNN) model with 96% accuracy. In a similar study, Haidar et al. (2017) built a CNN model capable of detecting OSA events with 75% accuracy using nasal flow signals. More recently, Yue et al. (2021) expanded upon this by using an ultra-deep CNN model with hundreds of layers to classify OSA severity based on raw nasal airflow signals and achieved an accuracy of 91%.

However, for EHR clinical data, DL methods have not been thoroughly studied so far. In Sect. 4, we will study the performance of various state of the art (SOTA) DL methods such as DANet, NODE or GATE, on OSA severity classifications and regressions with clinical data for the first time.

3.4 Data Preprocessing

There are wide ranges of data preprocessing techniques commonly used for OSA such as data imputation, imbalance correction, and feature selection.

3.4.1 Data Imputation

This is commonly used for EHR clinical data to deal with lots of missing information. The techniques used are mostly traditional methods. For example, median values are employed by Rodrigues Jr et al. (2020) for filling in missing data. However, the effects of these techniques on the overall prediction performance have not been studied thoroughly before. In Sect. 4.6, we study 16 different data imputation techniques for OSA severity predictions with our large-scale OSFP dataset.

3.4.2 Imbalance Corrections

All studied datasets in the literature are severely biased toward high OSA severities. For instance, 71.25% patients have $AHI \geq 30$ in Mencar et al. (2020). This makes learning models biased toward severe samples, thus reducing overall accuracy. Hence, many existing works employ different data imbalance correction techniques to balance class labels in the training data before feeding them to models (e.g., Rodrigues Jr et al. (2020)). However, the techniques used are limited to a few traditional methods, such as SMOTE (Chawla et al., 2002) and ADASYN (He et al., 2008). We will perform comprehensive studies of these methods and some advanced methods such as Borderline-SMOTE (He & Garcia, 2009) in combination with different feature selection methods and many learning models in Sect. 4.5.

3.4.3 Feature Selection

These are the most widely used techniques in existing works. However, their roles and effects have not been properly studied. To address this, we will also perform a comprehensive study on their effects in Sect. 4.4.

Depending on the kind of data, different features can be extracted. However, time-domain, frequency-domain, non-linear, and statistical features are the most common. For example, in Camcı et al. (2017), mean, variance, and range for accelerometer data and noise level are extracted and used with KNN and CART for apnea detection. For EHR clinical data in particular, features are commonly selected via statistical tests such as Chi-squared (Mencar et al., 2020; Ustun et al., 2016) or via feature rankings of common classifiers such as Information gain or gain ratio (Mencar et al., 2020). However, we will show that surprisingly none of these methods help to adequately improve prediction accuracy overall in Sect. 4.4. Despite this, reducing the number of features can still be useful to create simpler and more generalisable models for clinical use.

3.5 Phenotyping Problems

Different to the above mentioned supervised methods, unsupervised ML can be used to identify phenotypic clusters based on patient information. This could be more beneficial than the simple use of AHI to diagnose OSA as AHI does not take into consideration the heterogeneity of the disease. For example, a person may have a lower AHI yet suffer from excessive daytime sleepiness and be at high cardiovascular risk (Gottlieb & Punjabi, 2020; Appleton et al., 2018; Ahbab et al., 2013; Young et al., 2005; Javaheri et al., 2017). Under current guidelines, this person would not be recommended for PAP treatment for OSA despite its potential benefits (Kapur et al., 2017). Generally, cluster analysis pipelines involve gathering patient information including demographic variables, physical examination results, clinical history, co-morbidities, and/or PSG results; reducing the numbers of features using the Chi-squared test (Ye et al., 2014; Sobek, 1999) or PCA (Lacedonia et al., 2016; Vavougiotis et al., 2016); and clustering followed by cluster comparisons using Chi-squared, ANOVA, or Kruskal–Wallis tests.

For example, Ye et al. (2014) is generally cited as one of the first studies to use cluster analysis on OSA data in which 23 variables were used to generate 3 OSA clusters using latent class analysis for clustering. Several subsequent studies have also resulted in 3 (Lacedonia et al., 2016; Ferreira-Santos & Rodrigues, 2018) or 6 (Bailly et al., 2016; Vavougiotis et al., 2016) clusters. Both categorical (presence/absence of symptoms and comorbidities) and numerical (e.g., age, BMI, BP) data can be taken into consideration during cluster analysis. Ferreira-Santos and Rodrigues (2018) was the first to use k-modes categorical clustering while Vavougiotis et al. (2016) used PCA for categorical clustering. In Bailly et al. (2016), the largest-scale cluster analysis and included over 18,000 OSA patients, multiple correspondence analysis was performed on categorical clinical information, and ascending hierarchical clustering analysis was performed using the individual coordinates. Six OSA clusters, distinguished by age, weight, degree of symptoms, and number of comorbidities, was reported.

A major limitation of traditional cluster analysis is that cluster features must be manually identified, and the number of clusters must be manually determined either before (e.g. K-means) or after (e.g. hierarchical clustering) clustering. Ma et al. (2021) attempt to circumvent this by combining supervised and unsupervised ML. Their study included over 2,000 OSA patients and identified 6 clusters based on their multidimensional PSG data. PCA was used to reduce the number of features in consideration from 43 to 8, and the Dirichlet process Gaussian mixture model, which automatically determines the optimal

number of clusters, was used for clustering. Key features were identified via a supervised survival prediction model using RSF. The model was promising in that it was able to differentiate between OSA phenotypes; however, only numeric features were included, and the sample population lacked diversity.

Along with the above-mentioned works, references for current attempts to apply unsupervised clustering techniques can be found in Table 5.

3.6 Summarizations

3.6.1 Clinical vs. Non-clinical Data?

As mentioned above, existing works on OSA predictions rely on 3 main kinds of data: PSG-linked signals, Non-PSG signals, and Clinical data. While many signal-based models can classify OSA severity with very high accuracy (e.g. more than 90%), they require data which are hard to collect due to special device requirements (e.g., IR-UWB radar (Javaid et al., 2015) or 3D scanner (Islam et al., 2018)) and/or time consuming and involved procedures (e.g. ECG (De Chazal et al., 2020)). Hence, using routine and easily measured features, such as demographics (age, sex, etc.), measurements (BP, BMI, etc.), comorbidities, symptoms (snoring, sleepiness, etc.), and responses to questionnaires, has become an attractive target for many other research works (e.g., Wu et al. (2016); Mencar et al. (2020); Huang et al. (2020a)). Hence, they are the main target of our study thought the rest of this paper.

3.6.2 Current Situations and Limitations

Despite the efforts mentioned above on clinical data, many questions remain open.

First, existing studies are limited to small and selective datasets with 150 to 6,875 samples, as summarised in Table 4. As a result, the models used will be biased toward specific cohorts, reducing their generalisability. Patients with severe OSA are overrepresented in most datasets in the literature - 71.25% patients have $AHI \geq 30$ in Mencar et al. (2020), for example. Similar imbalances can also be seen for various demographics. Only 26% of individuals in Rodrigues Jr et al. (2020), 29.22% of individuals in Han and Oh (2023), and 24% of individuals in Mencar et al. (2020) are female. Compared to the slightly more balanced 41.8% seen in Ustun et al. (2016), these models will put less emphasis on using sex as a factor, potentially missing valuable diagnostic information. Additionally, the Wisconsin Sleep Study was conducted on primarily Caucasian participants, the research in Kim et al. (2021) was limited to the Korean population, and the dataset used in Wu et al. (2016) has a mean age of 41.7 ± 12.2 years, potentially missing out on the older populations in which OSA is frequently observed.

Second, most existing works use small sets of features, possibly missing many relevant OSA risk factors that could improve accuracy. For example, Han and Oh (2023) only use 17 features including demographics, body measurements, sleep questionnaires, and self-reported parameters, while comorbidities and symptoms, which are highly relevant to OSA (Pinto et al., 2016), are missed. Different studies also report different optimal subsets of features for their classifiers. Sex, for example, is an important feature in Mencar et al. (2020) but not in Rodrigues Jr et al. (2020). Taken together, these examples suggest the optimal combinations of prediction variables have yet to be identified.

Table 4 Supervised learning in OSA detection using clinical/non-signal data

Classifier (best)	#Subj.	Feature Set	Prob.	Acc.	Sens.	Spec.	RMSE
Neural Fuzzy Evaluation System (Wu et al., 2016)	150	age, BMI, ESS, waist circu., neck circu., 3 average differences of BP before going to sleep & early in the morning (Diff_S, Diff_D, Diff_M)	bin	75.60%	77.20%	75.00%	16.61
SLIM (Ustun et al., 2016)	1,922	sex, age, BMI, diabetes, hypertension, smoker	bin	—	64.2%	77.0%	—
SVM (Huang et al., 2020a)	6,875	age, waist circu., neck circu., snoring, sleep onset latency, witnessed apnea	bin	—	75.18%	74.71%	—
RF (Mencar et al., 2020)	313	BMI, ESS, reported apnea, diabetes, FEV1%	all	44.10%	44.10%	44.10% (Prec.)	22.17
Extra Trees (Rodrigues Jr et al., 2020)	1,042	age, nocturia frequency, BMI, depression score, neck circu., hip measurement, diastolic BP, smoker	all	—	64.0%	83.0%	16.25
NB (Keshavarz et al., 2020)	231	severe snore, nocturia, awakening due to the sound of snoring, witnessed snore, witnessed apnea, back pain, restless sleep, BMI	bin	—	78.95%	59.49%	—
SVM (Ramesh et al., 2021)	1,479	waist-to-height ratio, waist circu., neck circu., BMI, lipid accumulation product, excessive daytime sleepiness, daily snoring frequency & snoring volume	bin	68.06%	88.76%	40.74%	—
SVM (Kim et al., 2021)	279	hypertension, waist circu., length between the subnasale and stomion, snoring from the BQ, loudness of snoring from the BQ, frequency of falling asleep, FSS total score	bin	—	78.69%	73.91%	—
LightGBM (Han & Oh, 2023)	4,014	age, sex, BMI, abdominal circu., head, hip neck circu., ESS, ISI, K-BDI-II, PSQI, SSS scores, sleep hours, hypnotics consumptions	bin	91.0%	90.8%	70.1%	—

#Subj.: Number of Subjects, Number of Classes, Acc.: Accuracy, Sens.: Sensitivity, Spec.: Specificity, Prec.: Precision, RMSE: Root Mean Square Error, circu.: circumference. Prob.: Problem Type, whether it is a binary classes problem (bin) or an all classes problem (all)

Third, the effects of these data imbalances have not been carefully studied in the literature. For example, accuracy, which is sensitive to data imbalance, is a commonly reported metric to evaluate classifiers in many pieces of research (Mencar et al., 2020; Kim et al., 2021; Han & Oh, 2023) but it is also a misleading metric when class labels are highly imbalance as we will show in Sect. 4.3. Only a few works cope with this problem (Rodrigues Jr et al., 2020; Kim et al., 2021; Mencar et al., 2020) via some traditional data imbalance correction techniques like SMOTE (Chawla et al., 2002) or ADASYN (He et al., 2008).

Fourth, compared to using PSG signals, the reported accuracy of clinical data-based predictions is consistently much lower and inconsistent. For instance, Ustun et al. (2016) report an accuracy of only 44.1% for categorising individuals into the 4 main OSA severity groups {None, Mild, Moderate, Severe}. Similarly, Mencar et al. (2020) obtain an F1 score of 0.41, while a significantly different F1 score of 0.68 is reported by Rodrigues Jr et al. (2020). The prediction scenarios are also varied, including predicting distinct OSA severity classes (Ustun et al., 2016; Mencar et al., 2020; Rodrigues Jr et al., 2020), predicting binary classes

Table 5 Unsupervised learning in OSA study

Methodology	Co.	Im.	Ph.	R.F.	Sy.	Tr.
Cluster analysis	(Van Overstraeten et al., 2021; Lacedonia et al., 2016)	(Van Overstraeten et al., 2021; Lacedonia et al., 2016)	(Tsuchiya et al., 1992; Van Overstraeten et al., 2021; Lacedonia et al., 2016)		(Van Overstraeten et al., 2021)	(Van Overstraeten et al., 2021)
Latent class cluster analysis	(Ye et al., 2014; Keenan et al., 2018; Schütz et al., 2019; Bailly et al., 2021)		(Schütz et al., 2019; Bailly et al., 2021)		(Ye et al., 2014; Keenan et al., 2018; Cheng et al., 2023)	(Bailly et al., 2021)
Hierarchical clustering	(Bailly et al., 2016; Ferreira-Santos & Rodrigues, 2021)		(Bailly et al., 2016; Ferreira-Santos & Rodrigues, 2021; Ghorvei et al., 2025)	(Bailly et al., 2016)	(Bailly et al., 2016)	(Sutherland et al., 2021)
Two-step clustering	(Vavougiou et al., 2016)		(Vavougiou et al., 2016; Venkatnarayan et al., 2023)			
Evolutionary active constrained clustering			(Mai et al., 2018)			
Gaussian mixture model	(Ma et al., 2021)		(Ma et al., 2021; Ghorvei et al., 2025)			
K-mean clustering	(Razjouyan et al., 2024)	(Kim et al., 2020)	(An et al., 2020; Kim et al., 2020; Zhang et al., 2022; Ferreira-Santos & Rodrigues, 2023; Razjouyan et al., 2024; Ghorvei et al., 2025)			
Fuzzy C-mean clustering			(Ghorvei et al., 2025)			

Co.: Comorbidities, Im.: Impacts, Ph.: Phenotypes, R.F.: Risk Factors, Sy.: Symptoms, Tr.: Treatments

via different AHI thresholds (Kim et al., 2021), and predicting AHI values directly (Mencar et al., 2020). However, there is currently no comprehensive overview that consolidates these diverse prediction scenarios.

Fifth, though DL has been successfully proven a very effective approach for OSA signal data (see Tables 2 and 3), it has not been thoroughly studied in clinical data so far (perhaps) due to very small data sizes.

4 Part II: Comparative Experiment

In this Comparative Experiment Part, we aim to provide an in-depth study of OSA severity prediction using clinical data that can address the knowledge gaps and limitations discussed above. In order to achieve this, we focus on answering the following main questions: (i) how do different classifiers, including unexplored DL ones, perform when facing much

larger and more diverse datasets across a wide range of prediction scenarios and evaluation metrics?; (ii) how do different feature sets and data imbalances affect overall prediction accuracy?; and (iii) can we increase prediction accuracy by having more data?

4.1 Dataset

4.1.1 The OSFP Dataset

We conduct our studies on the OSFP dataset, which is collected by “*Observatoire Sommeil de la Fédération de Pneumologie*” (OSFP). OSFP contains 190,537 records that track the progress of consecutive visits for more than 110,000 patients. Severe OSA cases are more highly represented in the dataset, perhaps due to the fact that AHI is not generally measured unless OSA symptoms are severe (Sateia, 2014).

4.1.2 Data Preprocessing

Since patient information is recorded each time they come to the hospital/clinic, a patient has multiple records in the dataset; initially, we remove all records except the first for each patient. We only consider the first record because subsequent records may contain information affected by OSA treatment, which could lead to an inaccurate setup for our analysis. All patients without AHI values are also removed from the data since AHI is used to calculate OSA severity as explained below (Young et al., 2004; Gottlieb & Punjabi, 2020). The OSFP data contains a wide range of different features (attributes). Since the goal is to classify OSA severity based on easy-to-collect diagnosis data, we exclude all attributes that are linked to PSG, treatments, and OSA types. The remaining features are grouped into 5 categories: demographics, measurements, comorbidities, symptoms, and questionnaires. These are listed in Table 6 (continuous features) and Table 7 (binary features) in Appendices. At the end, we have a very large dataset with 80,390 patients and 51 different features.

4.1.3 Missing Data

Among 51 features in the pre-processed data set, 39 contained missing values. Features with more than 5% missing values are Diastolic BP (53.77%), Systolic BP (53.48%), Cervical perimeter (47%), Abdominal perimeter (46.32%), Depression scale (24.78%), Pichot scale (23.91%), Near miss accident (15.94%), Maxillofacial profile (15.64%), Epworth scale (11.81%), and Snoring (8.83%). Though many data imputation techniques can be used to fill in missing data, e.g. MICE (Van Buuren & Groothuis-Oudshoorn, 2011), they can alter the real distribution of data, thus affecting classification accuracy (Farhangfar et al., 2008). Hence, we use a subset of the data with 21,818 patients without any missing data for any features as the main data for our study in Sects. 4.3 to 4.8.1. However, we also study the full dataset with 80,390 patients in Sect. 4.8.1.

4.1.4 Data Labeling

Since we do not need the patient id column, we have 49 feature columns and 1 column for the AHI value, which is used for labeling patients based on their OSA severity levels, for our

studies. Following previous OSA research (Young et al., 2004; Gottlieb & Punjabi, 2020; Wu et al., 2016; Huang et al., 2020a), we focus on two main OSA classification settings:

- *The all class classification:* We divide OSA into 4 levels of severity: Severity 0 (i.e., *None*) with $AHI < 5$, Severity 1 (i.e., *Mild*) with $5 \leq AHI < 15$, Severity 2 (i.e., *Moderate*) with $15 \leq AHI < 30$, and Severity 3 (i.e., *Severe*) with $AHI \geq 30$. The percentage distribution of OSA labels from 0–3 was 14, 22, 22, and 42, respectively.
- *The binary classification:* Each binary task defines labels by an AHI threshold α such that if $AHI < \alpha$, we label patient as 0, if $AHI \geq \alpha$, we label patient as 1. We use 5, 15, 30 as three thresholds for 3 tasks.

4.1.5 Data Characteristics

Tables 6 and 7 in Appendices summarise the statistics of our cleaned and processed data (with missing values removed), presented for the entire cohort and for each severity class separately. Our dataset covers a wide range of ages from 18 to 111 with a reasonable ratio between men and women (62.06% vs 37.92%), which aligns with other research (Quintana-Gallego et al., 2004). The averaged values (for continuous features) and the percentages of positive values (for binary features) often monotonically increase or decrease across severity classes. For example, the average cervical perimeters are 37.88, 38.86, 39.65, and 41.92 (cm) for the *None*, *Mild*, *Moderate* and *Severe* classes, respectively, which is consistent with the significant correlation between OSA severity and cervical perimeters reported in other works (Ramesh et al., 2021). The percentages of patients with asthma are 14.40, 12.97, 12.62 to 9.33 for the four above classes, respectively. This decreasing trend has also been reported in Wang et al. (2023). However, the ranges of values for different classes are very similar. Cervical perimeter ranges are [25,60], [28,60], [26,60], and [28,60] for classes 0–3, highlighting the heterogeneity of the disorder.

4.1.6 What Makes OSFP Data Unique?

After preprocessing, the OFSP dataset includes 80,390 patients, with 21,818 having no missing data. This is 10.29 times larger than the largest dataset used in the literature ($n = 7,811$) and 3.17 times larger than the same dataset without missing values ($n = 6,875$) (Huang et al., 2020a). OFSP has 50 features that cover all 5 main categories commonly used for OSA predictions. Only the dataset from Ramesh et al. (2021) has a higher number of features (56), but it only has 1,479 samples (14.8x smaller), while the second largest number of features is 32 from Huang et al. (2020a).

Moreover, OSFP is highly diverse in comparison to other existing datasets. For instance, OSFP contains patients with ages ranging from 18.03 to 111.87 with a mean value of 53.26 ± 13.46 , which is more representative of the older populations in which OSA is commonly observed and more broadly distributed than the mean age of 41.7 ± 12.2 reported in Wu et al. (2016). The percentage of female patients in OSFP is 37.94%, which is much higher than all 12%, 24%, 24%, 29.22% and 26% of previous studies (Wu et al., 2016; Huang et al., 2020a; Mencar et al., 2020; Han & Oh, 2023; Rodrigues Jr et al., 2020). Only a few datasets such as the dataset from Ramesh et al. (2021), with 46.8% female, and SHIP-Trend (Fietze et al., 2019), with 46% female, are more balanced than OSFP, but they are

limited to a smaller sample sizes of mainly 1,479 Caucasian patients and 1,208 patients in northeast Germany, respectively.

Overall, with a very large number of patients and attributes, OSFP appears more diverse and generalisable than other datasets previously used in the literature. Hence, conclusions drawn from it are also expected to be more generalisable and less biased.

4.2 Experimental Settings

Experiments were conducted on a Linux server with a 64-core CPU (Intel Xeon E7-4860 v2 - 2.6 GHz) and 32 GB RAM using Python 3.11.

4.2.1 Model Selections

We employ 39 different classification models, which are specifically designed for tabular data, for different OSA prediction scenarios:

- *27 ML models:* Nearest Centroid, Hist Gradient Boosting, Gradient Boosting, LGBM (Ke et al., 2017), MLP, Bagging, Random Forest, Stochastic Gradient Descent (SGD), XGBoost, Gaussian NB, AdaBoost, Linear Discriminant Analysis (LDA), Multinomial NB, Logistic Regression, CatBoost (Prokhorenkova et al., 2018), Extra Tree, Bernoulli NB, Quadratic Discriminal Analysis (QDA), Support Vector Machine (SVC), Linear SVC, Decision Tree, Ridge, KNN, Perceptron, Passive Aggressive, Radius Neighbours, and NuSVC.
- *12 DL models:* Gated Recurrent Unit (GRU), DANet (Chen et al., 2022), TabTransformer (Huang et al., 2020b), LSTM, GATE, CEM, Node (Borisov et al., 2022), RNN, Deep Belief Network (Hinton et al., 2006), TabNet (Arik & Pfister, 2021), GCN and CNN. Most of these have not been used in OSA research before (for clinical data only).

For the AHI prediction problem, we employ the same models but as regressors rather than classifiers. With a huge number of models studied, we hope that our study can provide a comprehensive picture to support other researchers to choose suitable models for their data.

4.2.2 Performance Metrics

For the classification problems, we employ 6 different evaluation metrics: Accuracy, Precision, Recall, F1, Balanced Accuracy, and G-Mean. While Accuracy, Precision, Recall, F1 are common metrics in OSA research, Balanced Accuracy and G-Mean (Kelleher et al., 2020) have not been used before. They are specifically designed to cope with data imbalance, a major challenge in OSA research. For the regression task, we employ common metrics such as RMSE and MAE.

4.2.3 Stability

All of the results are reported via 5-fold cross-validation with stratified sampling.

4.2.4 Hyper-parameter Optimisations

For all employed methods, we use grid-search to find optimal hyper-parameters. The detailed parameter spaces can be found in Tables 10 and 11 in Appendices for DL and ML models, respectively.

4.3 OSA Severity Classifications

We study the performance of each ML/DL method using two common frameworks from the literature - all classes prediction and binary prediction - in Figs. 3 and 4, respectively.

4.3.1 All Classes Prediction

Figure 3A shows the performance of all studied methods over 5 different folds and evaluation metrics (sorted and separated into ML and DL parts), using the full set of 49 features described in Sect. 4.1.1. For Acc, CatBoost was the best method among ML ones with an average Acc of 0.4690, followed by LDA, Logistic Regression, XGBoost, and Linear SVC with very small performance differences (from 0.4686 to 0.4642). With an average value of 0.4730 and fold values of {0.479, 0.463, 0.481, 0.474, 0.466}, NODE was the best performing method across both ML and DL categories.

A few key observations can be seen: (i) different methods behave differently with regard to evaluation metrics - NODE, GATE, LGBM, and DANet have the best respective Bacc, F1 and GM values, for example. However, this is a normal phenomenon as each metric captures a different meaning of the results, especially when class labels are not equally distributed; (ii) the top methods in both ML and DL groups differ very slightly in all performance metrics (though the best ones are usually DL). For example, LGBM, AdaBoost, Random Forest, GATE, CEM, and DANet have very close average F1 scores, ranging from 0.4230 to 0.4340; (iii) some ML methods (e.g., Perceptron, Passive Aggressive, and NuSVC) and DL ones (e.g., GCN and TabNet) have unstable performances, i.e., their metrics varied significantly over different folds (indicated via wider value range areas). Ensemble approaches (e.g., CatBoost, Random Forest, XGBoost, AdaBoost, Hist Gradient Boosting, Bagging, LGBM or Extra Tree) have consistent performances over all folds due to their multi-result aggregation (fusion) schema, while most DL methods show less stable results than them, especially on the GM metric; and (iv) overall performances are low for all metrics. Particularly, the best average values are 0.473, 0.393, 0.434, 0.333, 0.473, and 0.437 for Acc, Bacc, F1, GM, Recal, and Prec, respectively, indicating relatively weak classification abilities among all 4 classes compared to randomised guesses. We will study these key points in greater detail below.

Which methods perform consistently well over all metrics? To answer this, we rank all methods from 1–39 based on their performance on each metric separately (Fig. 3B). As we can see, the rank of each method varies considerably for all metrics (indicated by wider value ranges), consistent with the observations in (A). Among ML methods, LGBM had the best average ranking of 7.16 followed by LDA, AdaBoost and Random Forest with averaged rankings of 9.66, 10.33 and 10.83, respectively. Among DL approaches, DANet was the best one with averaged ranking of 4.16 followed by GATE, CEM and Node with corresponding values of 4.83, 5.16 and 8.83 resp, which are much better than LGBM and all

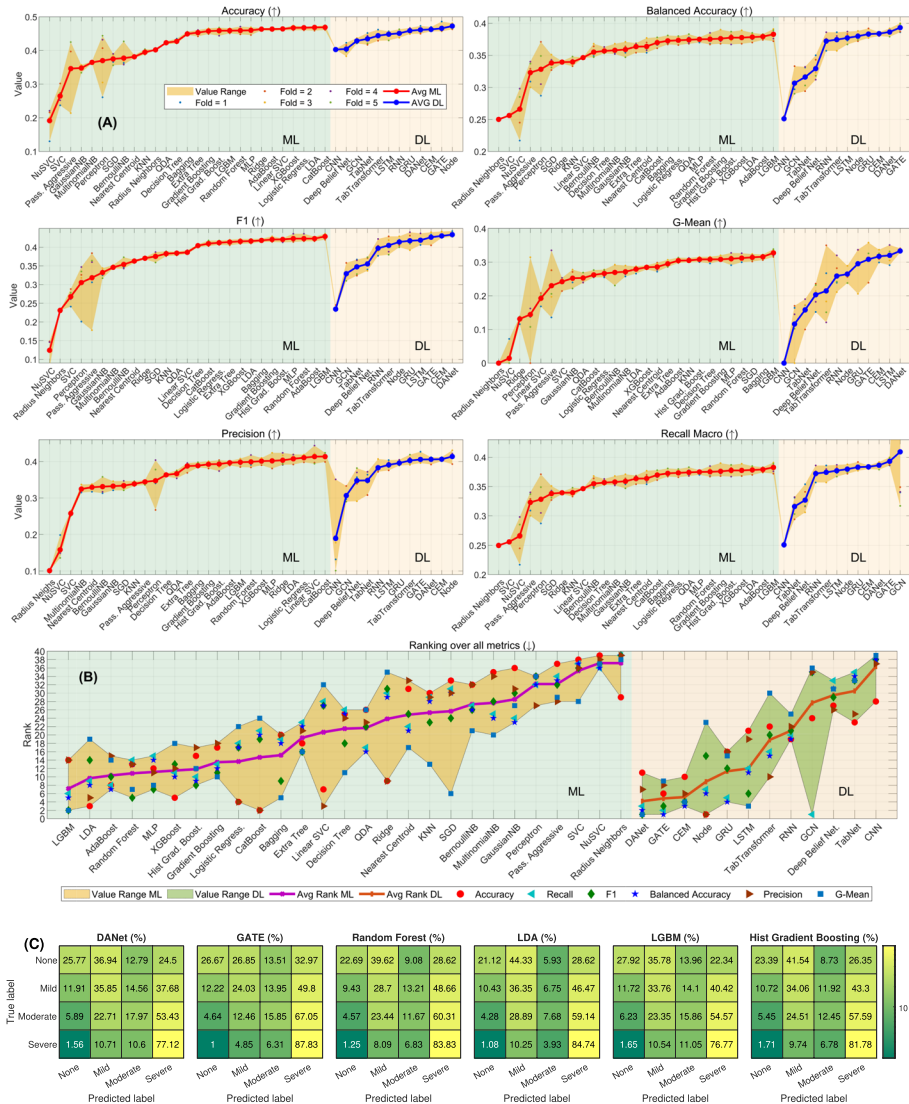


Fig. 3 Performances of studied ML/DL methods for the 4-classes classification problem using the full feature set. **A** Performance metrics (Accuracy, Balanced Accuracy, F1, G-Mean, Precision, Recall) for all methods over 5-folds (sorted and separated into ML and DL parts). **B** Performance rankings of all methods on all metrics (sorted and separated into ML and DL parts). **C** The confusion matrices for one fold of various methods. Best viewed in color

other ML techniques. Hence, DL would be a better choice than ML. However, as shown in (A), DL methods tend to show less stable performance than ML methods, especially boosting and bagging techniques.

Figure 3C shows the confusion matrices for some top-performing models. All of them can predict the severe class well with accuracies from 77.12% to 87.83%. However, for all other classes, the prediction accuracies are below 36.35%. Moreover, major parts of the

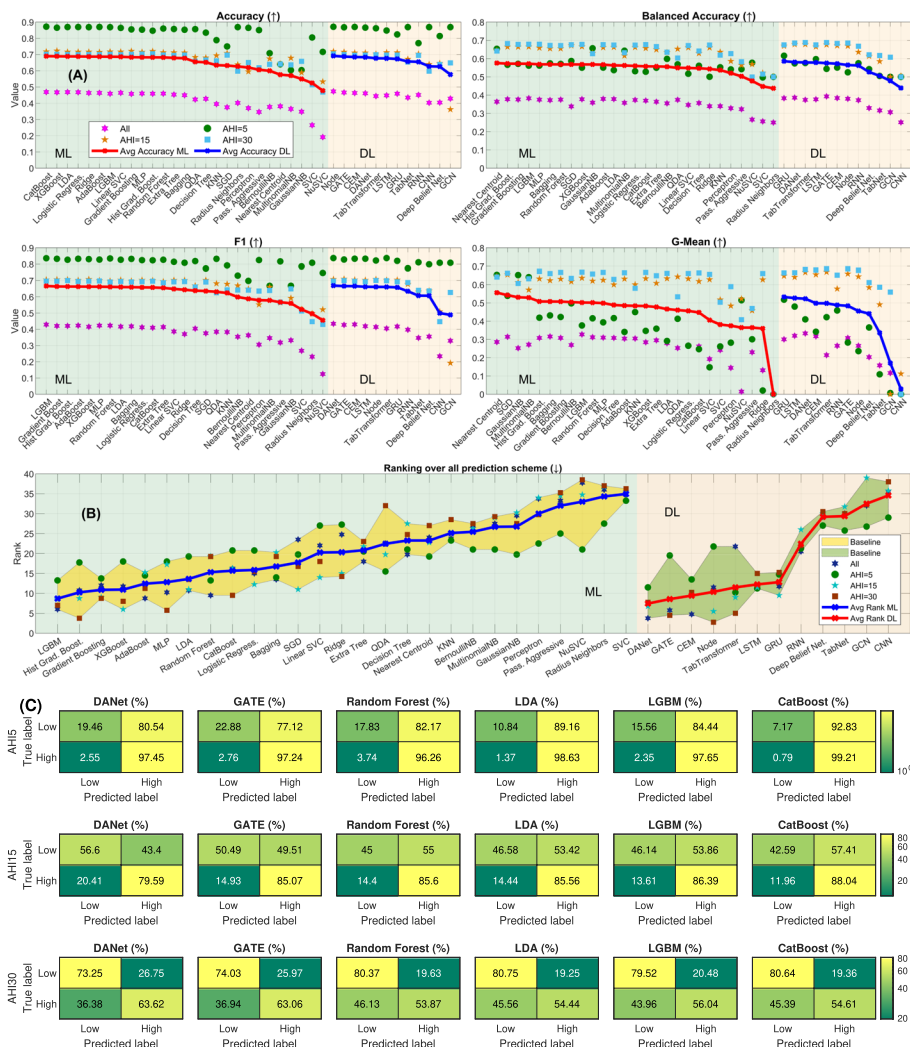


Fig. 4 Performance of all methods in the binary settings with different AHI thresholds in {5, 15, 30}. **A** Averaged performance metrics (Accuracy, Balanced Accuracy, F1, G-Mean) for all methods over all folds (sorted by averaged values of all metrics and separated into ML and DL parts). **B** Performance ranking of all methods over all metrics and prediction scenarios. **C** Confusion matrices for some top methods on one fold. Best viewed in color

mild and moderate classes are wrongly classified as severe cases (up to 49.8% and 67.05%, respectively), while the none (no OSA) class are typically classified into mild and (surprisingly) severe (up to 44.33 and 32.97%, respectively). One reason for these biased outcomes might be the uneven class distributions in our data, in which the severity class contains 42% samples, while the none, mild and moderate classes have 14%, 22% and 22%, respectively. This imbalance distribution causes learning methods to be biased towards the severity class (Chawla et al., 2002; He et al., 2008; He & Garcia, 2009), thus leading to uneven predic-

tion outcomes as also noted in other OSA research before (Khan et al., 2023; Álvarez et al., 2020; Rodrigues Jr et al., 2020).

4.3.2 Binary Classes Prediction

Fig. 4A shows the performance of all methods across different common prediction scenarios (all binary classes with AHI thresholds of 5, 15 and 30, i.e., we predict if AHI is above threshold (High) or below (Low)) (Ustun et al., 2016; Huang et al., 2020a; Rodrigues Jr et al., 2020). By reducing the total number of classes to be predicted, the classification task becomes easier, thus resulting in higher evaluation metrics compared to the 4-class case. For example, for 4-classes, LGBM has an average Bacc of 0.3828. But its average Bacc scores rise to 0.5616, 0.6648 and 0.6786 for AHI thresholds 5, 15, and 30, respectively. Similarly to the 4-class case, the best DL and ML methods have very similar performance for each scenario and metric. For example, LGBM has F1 scores of 0.8360, 0.7054 and 0.6968 for AHI thresholds 5, 15, and 30, respectively, while the corresponding scores for DANet are 0.8370, 0.7080, and 0.6970. Also, though DL models tend to have slightly better performance in many cases, they are less stable than ML methods, especially boosting tree approaches like LGBM or XGBoost. Moreover, different methods show different relative performances compared to others under different metrics, especially Bacc and GM, two class balance-aware measures. For example, Nearest Centroids is among the best methods in terms of Bacc (64.63%) and GM (64.63%) but is among the worst ones in terms of Acc (64.1%) and F1 scores (66.27%).

Which methods perform consistently well on all prediction scenarios? For each scenario, we rank all methods based on four main metrics (Acc, F1, Bacc, and GM) and report their average ranking across all binary prediction scenarios (Fig. 4B). DANet, GATE, CEM and Node, with average rankings of 7.43, 8.56, 9.43 and 10.37, are the best of all DL methods, while LGBM, Hist Gradient Boosting, Gradient Boosting and XGBoost are the best among ML ones with averaged rankings of 8.68, 10.25, 10.87, and 10.93, respectively. The results are consistent with the 4-class case in Fig. 3B, where DANet and LGBM are the two leading methods in DL and ML, respectively.

Figure 4C shows confusion matrices for different algorithms and binary prediction scenarios. For $AHI = 5$, only 19.46% (by DANet) of low samples are correctly predicted while most of them are incorrectly classified as high ones. For $AHI = 15$, the accuracy of the high class is reduced to around 80% but the low class is increased to more than 40%. However, this means that over half of the low class samples are still wrongly classified into the high class. For $AHI = 30$, up to 46.13% of the high class samples are wrongly put into the low class, while up to 26.75% low class samples go into the high one. Hence, the high Acc and F1 scores shown in (A) can be misleading by implying we have very good classification results when it is actually an artefact of high class-label imbalance. In particular, the low and high class ratios are 14:86, 36:64, and 58:42 for AHI thresholds 5, 15, and 30, respectively. Bacc and GM scores, which are designed to deal with data imbalance, seem to be better choices since they provide lower values than F1 and Acc in highly imbalanced cases ($AHI = \{5, 15\}$) while resulting in similar values to Acc and F1 in the more balanced case ($AHI = 30$). While data imbalances may be responsible for the less than satisfactory results in binary $AHI = \{5, 15\}$ cases and the 4-class case above, the binary $AHI = 30$ case tells us that classifying OSA severity is a difficult task even when we have large and

balanced data to train models with. Nevertheless, we will study the data imbalance problem in more depth in Sect. 4.5.

4.3.3 Summary

To summarise, the best DL methods (e.g., DANet, GATE, CEM and Node) perform slightly better than the best ML ones (mainly boosting and bagging techniques such as LGBM and XGBoost), possibly due to DL's ability to learn complex patterns from the data (as summarised in Fig. 5 using Wilcoxon signed rank test at 5% significance level across all prediction settings and evaluation metrics). However, the performance improvements are marginal and usually come at the expense of worse stability across different training–testing folds. Predicting binary classes is easier than predicting which group patients fall under in the 4-class problem, resulting in higher evaluation metrics.

However, overall prediction performance is (relatively) low and is biased toward the Severe class for all methods studied, mirroring the findings of some other works (Wu et al., 2016; Rodrigues Jr et al., 2020; Mencar et al., 2020). For example, for the all class task, an Acc of 0.441 is reported by Mencar et al. (2020) using a Random Forest model with 313 samples, whereas we acquire a slightly better score of 0.4594. In contrast, Rodrigues Jr et al. (2020) report an F1 of 0.66 on Extra Tree, significantly higher than both our Extra Tree F1 score of 0.4138 and our highest F1 score of 0.4336 with DANet. Han and Oh (2023) report over 90% accuracy for binary classification, which is much higher than our current results. This may be due in part to our dataset being larger and more heterogeneous, with a higher proportion of lower-severity patients.

4.4 Feature Selections

As mentioned in Sect. 3, different existing works employ different sets of input features for their OSA classifiers (Ustun et al., 2016; Rodrigues Jr et al., 2020; Keshavarz et al., 2020;

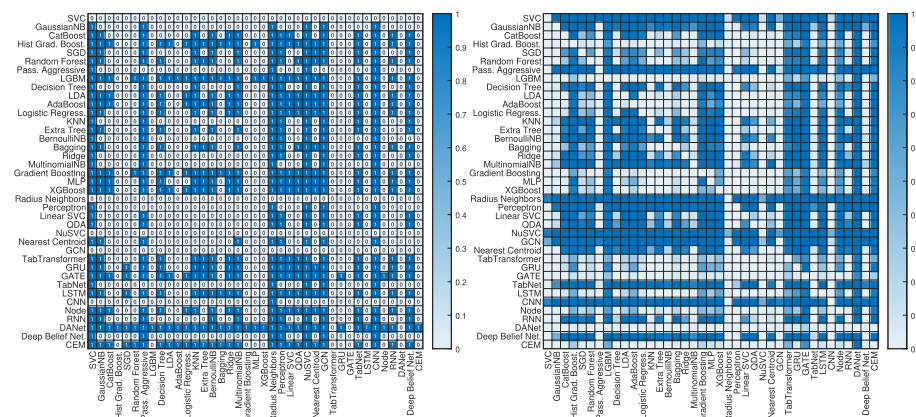


Fig. 5 The Wilcoxon signed rank test for each pair of methods (x, y) across all classification settings and evaluation metrics. (Left) Result of the hypothesis test h , where 1 mean that the row method x is better than the column method y (rejection of the null hypothesis at 5% significance level). (Right) The corresponding p -value of hypothesis test on the left side. For example, DANet is better than LightGBM with p -value of 0.0081. However, it is not statistical better than GATE and TabTransformer with p -values of 0.2356 and 0.1033, respectively. Best viewed in color

Kim et al., 2021), which usually contain around a dozen of manually selected features (see Table 4 for more details). There is no universally accepted feature set for the task, except for some common features like age or BMI. In this paper, we have a very large dataset with a huge set of 50 different features that cover all common relevant factors of OSA studied in the literature. Hence, it raises a interesting question: which feature subsets can help to improve the prediction accuracy? In this Section, we will study deeper the effects of using different subsets of features on the performance of all classifiers under the same 4-class and binary class prediction scenarios studied above in Sect. 4.3.

4.4.1 Feature Set Constructions

Table 8 presents 16 sets of features used in our study. They are organised into 3 different categories: characteristics, statistics, and literature.

- *Characteristics*: This category contains 5 groups representing different kinds of attributes: demographics (e.g., age or gender), measurements (e.g., BMI), comorbidities (e.g., diabetes or hypertension), symptoms (e.g., snoring or headache), and questionnaires (e.g., Epworth, Pichot and depression scales) as commonly found in other works (Bailly et al., 2016).
- *Statistics*: We leverage common statistical analyses (e.g., correlation and Chi-squared (Cochran, 1952)) as well as model-based feature selections (e.g., Random Forest and CatBoost feature ranking schema (Prokhorenkova et al., 2018) or Shapley Additive explanation (SHAP)-based feature importance selection (Wang et al., 2024)) to create highly relevant feature sets to OSA severity.
- *Literature*: We employ manually selected feature sets reported in previous studies - Wu et al. (2016), Mencar et al. (2020), Huang et al. (2020a), Ustun et al. (2016), and Rodrigues Jr et al. (2020).

4.4.2 All Classes Predictions

Figure 6A illustrates the performances of all ML/DL models included in the study using the 16 different sets of features described above. The results of the full feature set are used as baselines. All models are sorted based on their baseline performance. Each subset has different effects on different classifiers and evaluation metrics. Performance decreases significantly compared to the baselines in most cases. However, a suitable feature set can help to improve the performance of some weak models considerably. This can be seen for GaussianNB, where accuracy improves from 0.3478 with the full feature set to 0.4510 with the CatBoost-chosen one. Moreover, 15/16 sets (i.e., 93.75%) increase Acc scores for GaussianNB compared to the full feature usage. However, improvements are negligible for the top-performing models. For example, none of the feature sets actually improve the performance of DANet for any metric, and the Bacc score of Hist Gradient Boosting only increases very slightly from 0.4090 to 0.41 with the RF Permutation set. Moreover, the models that see performance gains from using subsets of features still fall short of the best models using the full feature set. For example, though GaussianNB can reach an Acc of 0.4510, this is still lower than the baseline Acc of 0.4690 from CatBoost.

Which subsets consistently perform better than others for all classifiers and metrics? For two subsets x and y , we count the number of times that using x leads to better performance than using y for all classifier-metric combinations (39 models and 4 metrics) (Fig. 6C). Among all subsets, CatBoost, RF Permutation, Correlations and SHAP RF are leading ones, while Comorbidities, Questionnaires and Symptoms are the worst. However, using all features is still the best choice. For example, it wins against RF Permutation in 69% of cases while losing in only 25% cases over $39 * 4 = 156$ classifier-metric cases. However, this

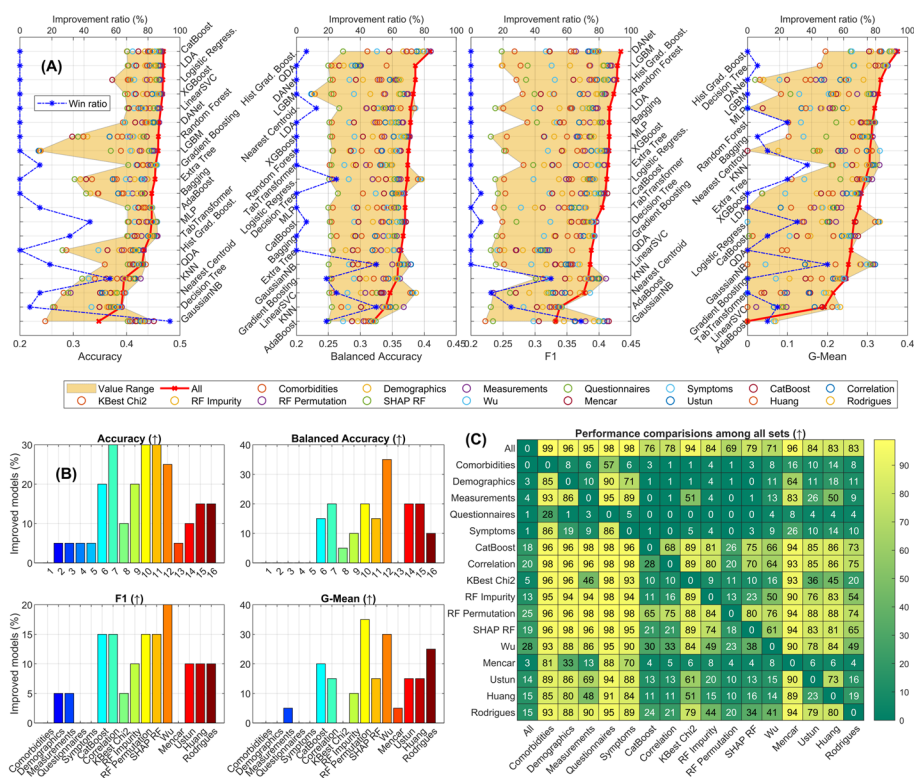


Fig. 6 Performances of various ML/DL models wrt. different feature sets. **A** Performance metrics (Accuracy, Balanced Accuracy, F1, G-Mean) for all methods and feature sets (sorted by performances on all features, denoted as *All*). The win ratio lines show the percentages of performance improvement cases over all 16 feature sets for each model compared to the full feature set (i.e. all possible features). **B** The numbers of improved models for different feature sets wrt. different evaluation metrics (in percentages over all classifiers). **C** Performance comparisons among all feature sets (aggregated over all combinations of classifiers and evaluation metrics). A number x at row i and column j denotes the feature set i is $x\%$ times better than the feature set j over all classifier/metric combinations (i.e., i has $x\%$ probability to have better results than j given an arbitrary classifier and a metric). E.g., CatBoost is better than Comorbidity in 96% testing cases while losing to Correlation in 28% testing cases. Best viewed in color

does not mean we should not use feature selections at all, since they still can help to improve the performance in specific cases as shown in Fig. 6A, B.

4.4.3 Binary Classes Predictions

The effects of all feature subsets on the binary classification scenarios (with AHI thresholds 5, 15, and 30) are presented in Fig. 7. We have exactly the same observations as in the 4-classes case above.

4.4.4 Summary

Overall, though good feature sets can help to significantly improve the performance of some weaker ML/DL models like KNN or GaussianNB, they have negligible effects on stronger ones like DANet or LGBM. This finding is consistent with previous studies that have found little difference in performance from using subsets (Ustun et al., 2016; Mencar et al., 2020; Rodrigues Jr et al., 2020; Huang et al., 2020a). However, this is not inherently bad. Maintaining performance with fewer features can reduce the risk of overfitting, reduce training times, and provide a simpler clinical tool that is more adaptable to situations where patient data is limited.

We also find statistical methods such as CatBoost, Correlation, or RF Permutation to be more effective than using qualitative categories such as symptoms or comorbidities when choosing suitable feature subsets to use (Fig. 6C). Given that many patients with OSA have few symptoms (Gottlieb & Punjabi, 2020), it is perhaps unsurprising that classifying OSA using symptoms alone leads to poorer performance in our study and in previous work (Ustun et al., 2016). Nevertheless, using all features remains the most effective method across all prediction settings, classification methods and evaluation metrics as indicated in Fig. 8 using the Wilcoxon signed rank test with 5% confidence level. For feature subsets, CatBoost, Correlation, and RF Permutation are statistically better than others.

4.5 Data Imbalance Corrections

In data classification, when the labels of training data are unevenly distributed, the learning algorithms will be biased toward the largest classes and therefore do not perform well when predicting the smaller ones (He & Garcia, 2009). This phenomenon, known as *imbalanced data*, is a common problem in OSA severity classification since all studied datasets suffer from label imbalance at different levels. The imbalance ratio (i.e., the ratio between the largest and smallest classes) in our OSFP data is around 3, while the ratios in previous studies (Rodrigues Jr et al., 2020; Huang et al., 2020a; Ustun et al., 2016) are 8.3, 4.7 and 3.3, respectively. To deal with this problem, data oversampling techniques are commonly used in OSA research, notably two SOTA methods: SMOTE (Mencar et al., 2020; Kim et al., 2021) and ADASYN (Rodrigues Jr et al., 2020). Hence, understanding how these techniques interact with different classifiers, feature selections, and prediction scenarios will be extremely useful for designing future methods. Besides SMOTE and ADASYN, we additionally study a few more advantaged data imbalance techniques in OSA: Borderline-SMOTE and Undersampling (He & Garcia, 2009).

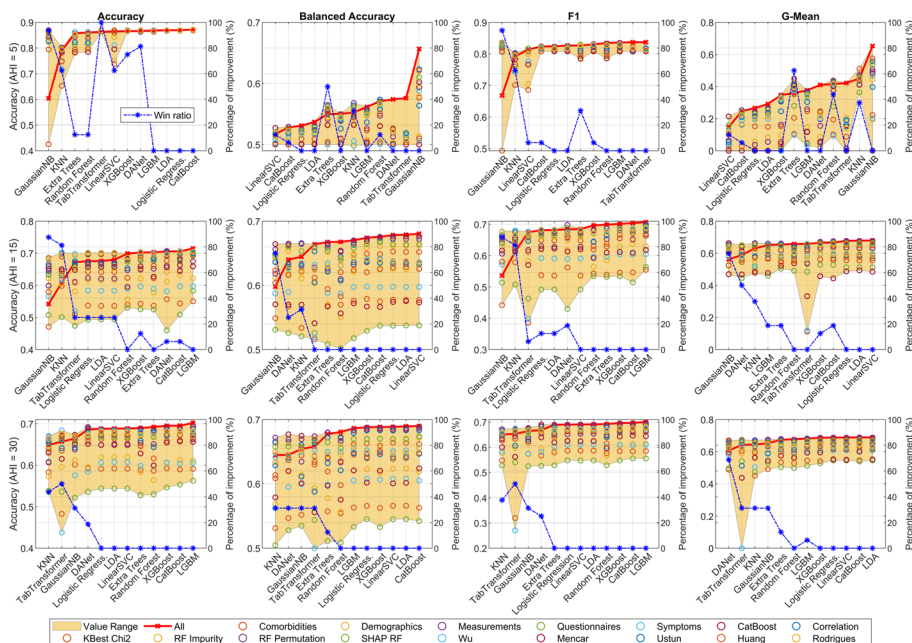


Fig. 7 Performances of various ML/DL models (Accuracy, Balanced Accuracy, F1, G-Mean) wrt. different feature sets under 3 binary prediction scenarios with AHI thresholds 5 (top), 15 (middle) and 30 (bottom). All methods are sorted via their performances on the full feature set (denoted as *All*). The win ratio lines show the percentages of feature sets that lead to performance improvements over the full feature set. Best viewed in color

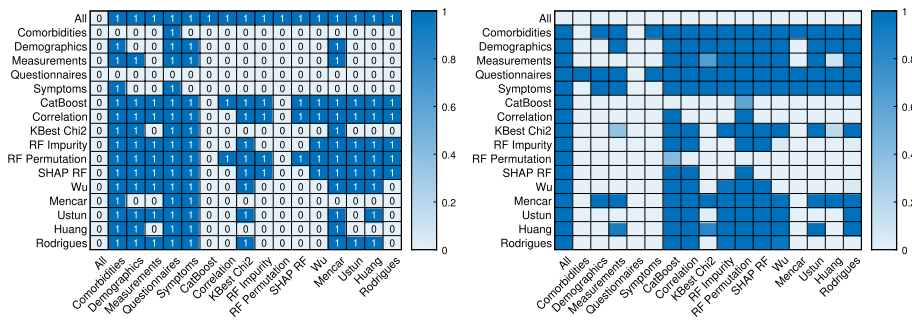


Fig. 8 The Wilcoxon signed rank test for each pair of feature sets (x, y) across all classification settings, learning models and evaluation metrics. (Left) Result of the hypothesis test h , where 1 means that the row set x is better than the column set y (rejection of the null hypothesis at 5% significance level). (Right) The corresponding p -value of hypothesis test on the left side. For example, *All* is better than *CatBoost* with p -value of 0.000035. However, *CatBoost* is not statistically better than *RF Permutation* with p -values of 0.5747. Best viewed in color

4.5.1 All Classes Predictions

Different imbalance techniques can lead to significant changes (both increases and decreases) in the performance of many classifiers, especially for weak models and imbalance metrics such as Bacc and G-Mean (Fig. 9A). For example, the Bacc score of XGBoost improves from a baseline of 0.3776 to 0.4030, 0.4028, 0.3990, and 0.4064 with ADASYN, SMOTE, Borderline, and Undersampling, respectively. However, these improvements rarely beat the best-performing models. For example, while some models show substantial improvements in GM score after imbalance correction (e.g., AdaBoost, CatBoost), they still fall short of the 0.3764 GM score of the unmodified Hist Gradient Boosting model. However, Undersampling does slightly increase the Hist Gradient Boosting GM score to 0.3774. Hence, it is important to explore various combinations of both classifiers and imbalance correction techniques when trying to improve performance.

When it comes to comparing techniques, the best method to choose depends on which metric one wishes to optimise. In particular, SMOTE was the most effective for improving accuracy, SMOTE and Borderline were the most effective for improving F1, and undersampling was the most effective for improving G-Mean (Fig. 9B). Additionally, borderline correction was the most consistent technique, displaying better performance more than 50% of the time when compared to other method (Fig. 9C).

Compared to the original full feature data shown in Fig. 3C, some small improvements can be seen from using SMOTE (Fig. 9D). For example, Random Forest can classify *None*, *Mild*, and *Moderate* classes slightly better, but the accuracy for the *Severe* class decreases considerably. Nevertheless, the same problem remains. Large portions of other classes are wrongly classified as *Severe*. This is why improvements can be seen in many cases but the overall performance metrics are still low as also seen in Sect. 4.3. In other words, balancing data labels alone is not a magic solution for improving accuracy. Again, these results do not mean we should not use feature subsets and/or imbalance corrections, but they should be used with caution.

What happens when we combine both feature subsets and imbalance techniques? In many cases, performance goes down significantly compared to using neither technique or only using imbalance correction (Fig. 9E, F). Improvements can be seen in a few cases, particularly for weaker models on balanced accuracy and G-mean. This is consistent with the results from Sect. 4.4. Given an arbitrary model, using all features without imbalance correction is still the most stable solution in general (Fig. 9F). However, other combinations such as (All, SMOTE), (All, ADASYN), (Catboost, SMOTE), or (RF Permutation) perform very similarly.

4.5.2 Binary Classes Predictions

Fig. 10 studies the performances of all classifiers and imbalance methods for all three binary classification scenarios ($AHI = \{5, 15, 30\}$). The same results are observed as in the 4-class case studied above.

4.5.3 Summary

Compared to feature selection (Sect. 4.4), data imbalance methods have stronger effects on performance. Although performance varies considerably with different imbalance methods

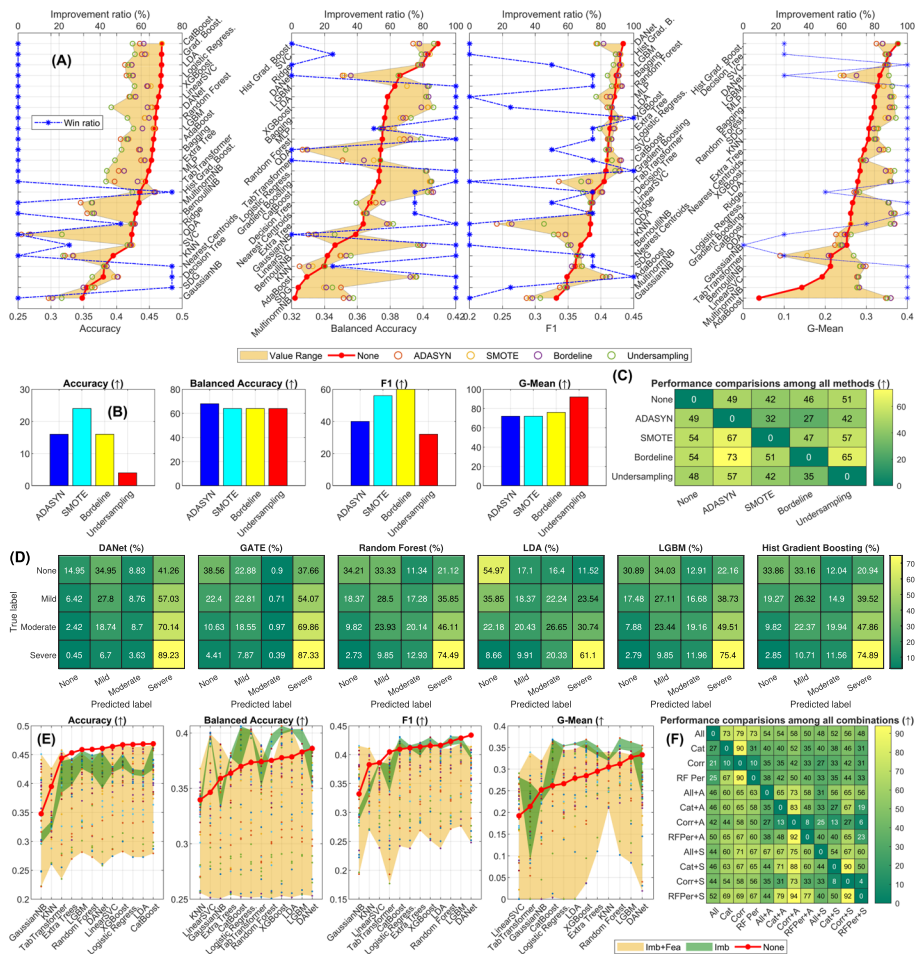


Fig. 9 Effects of imbalance data correction techniques and feature selections on various classifiers. **A** Performance metrics of different models wrt. different imbalance data correction methods using the full feature set (sorted by their performance without imbalance corrections, denoted as *None*). **B** The numbers of improvement cases (in percentages) for each imbalance techniques over all studied models for each evaluation metric. **C** Performance comparisons among all imbalance techniques. **D** Confusion matrices for various classifiers with SMOTE. **E** Performance metrics of various classifiers combined with different imbalance techniques and feature sets (Imb+Fea) and imbalance methods only (Imb). Each dot presents one result. **F** Performance comparisons among various imbalance techniques and feature sets (aggregated over all combinations of classifiers and metrics). Best viewed in color

and classifiers, SMOTE and Borderline consistently perform well on many classifier and metric combinations as validated via the Wilcoxon signed rank test presented in Fig. 11.

When used with feature selection, they can lead to further performance improvements in some cases but can also lead to significantly performance reduction if not used carefully. Among them, combinations of (Correlation, ADASYN) and (RF Permutation, ADASYN) perform consistently better than others on arbitrary classifier-metric selections. However, none of them is a game changing one to improve the classification accuracy, especially for

the 4-class problem. Our findings are in line with previous works. E.g., Mencar et al. (2020) reports insufficient performance improvements for the 4-class problem with SMOTE and Random Forest. Rodrigues Jr et al. (2020) observes significant improvements in F1 scores of minority classes but also considerably F1 reductions of the *severe* class with ADASYN and Naive-Bayes.

4.6 Do We Need More Data?

4.6.1 All Class Predictions

Figure 12 shows the results when using the full preprocessed OSFP data with 80 K patients. Since many patient records contain missing data, we use imputation methods to fill them before using. In Fig. 12A, effects of 16 different imputation techniques (c.f. Appendices for details) on the performance of various classifiers (e.g. Random Forest and LGBM) are presented. For all metrics, the result changes very slightly indicating small effects. So, we choose Median Const as the main filler method. In Fig. 12B, we show the performances of various classifiers with/without data imbalance corrections using the full feature set for both small and large OSFP datasets. Overall, using larger dataset does not help to improve the performance.

In order to assess the impact of a larger dataset, we re-ran our analysis using the full OSFP dataset of 80 thousand patients. Since many patient records contain missing data,

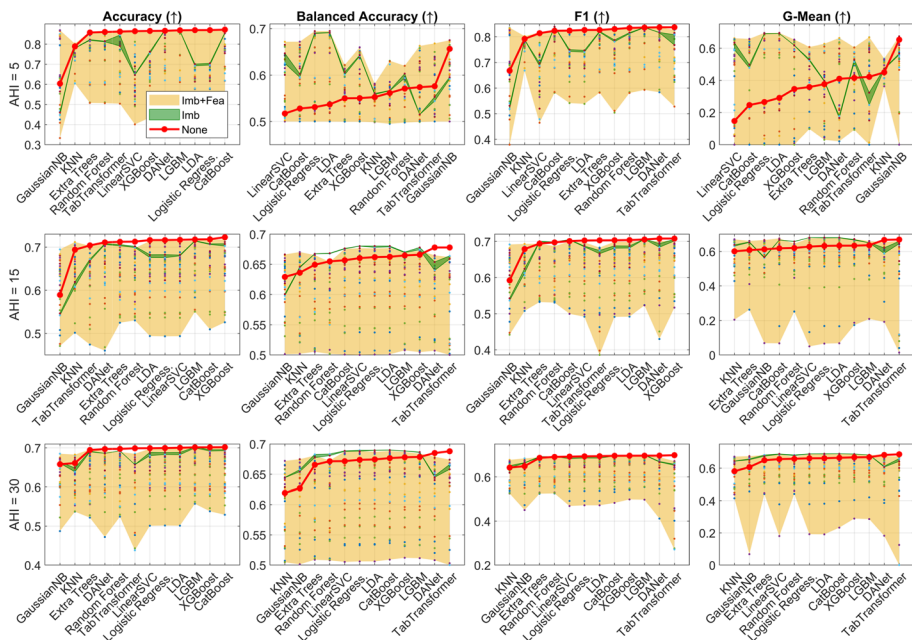


Fig. 10 Performance metrics (Accuracy, Balanced Accuracy, F1, G-Mean) of various classifiers when being combined with different imbalance techniques and feature sets (Imb+Fea) and imbalance methods only (Imb) for 3 binary classification scenarios ($AHI = \{5, 15, 30\}$). Each dot presents one result. All classifiers are sorted base on their performance on raw data with all features, denoted as *None*. Best viewed in color

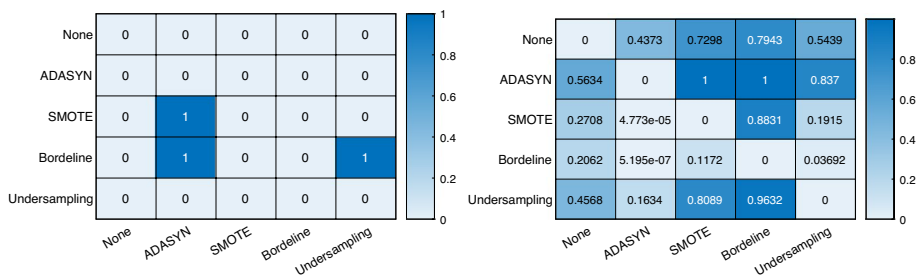


Fig. 11 The Wilcoxon signed rank test for each pair of imbalance methods (x, y) across all classification settings, learning models and evaluation metrics. (Left) Result of the hypothesis test h , where 1 mean that the row method x is better than the column method y (rejection of the null hypothesis at 5% significance level). (Right) The corresponding p -value of hypothesis test on the left side. For example, SMOTE is better than ADASYN with p -value of 0.000047. Best viewed in color

we tested 16 different imputation techniques, finding only small differences between them (Fig. 12A). Similarly, after selecting Median Const as a main filler method, there were no improvements in overall performance from using the larger dataset (Fig. 12B).

4.6.2 Binary Predictions

The same results are observed for the binary problems in Fig. 13. No improvement is observed with the larger dataset overall. The performance even becomes worse with AHI=15 and AHI=30.

4.6.3 Summary

While having much larger training data is expected to improve overall performance, especially for DL classifiers, this does not apply in our case. Perhaps, imputation creates a noisier dataset that decreases performances.

4.7 Direct AHI Prediction

Instead of classifying OSA severity, a few works aim to predict AHI values directly (e.g., Rodrigues Jr et al. (2020); Mencar et al. (2020); Wu et al. (2016)). Figure 14A shows the performances of various regressors on small and large datasets with 22 K and 80 K samples, respectively. The best RMSE is 18.8 of LGBM, which is consistent with reported results from previous studies (Rodrigues Jr et al., 2020; Mencar et al., 2020). When increasing sample size, performance decreases in most cases. Again, this may be due to high amounts of missing data. In Fig. 14, we show an example of the actual and predicted AHI values on a test set of DANet. Around 35.3% predicted values are larger than 30, while 63.86% are within the range of [5,30]. These findings are in line with the classification results studied above. Overall, with real-world AHI ranges from [0, 120], prediction accuracy is quite poor.

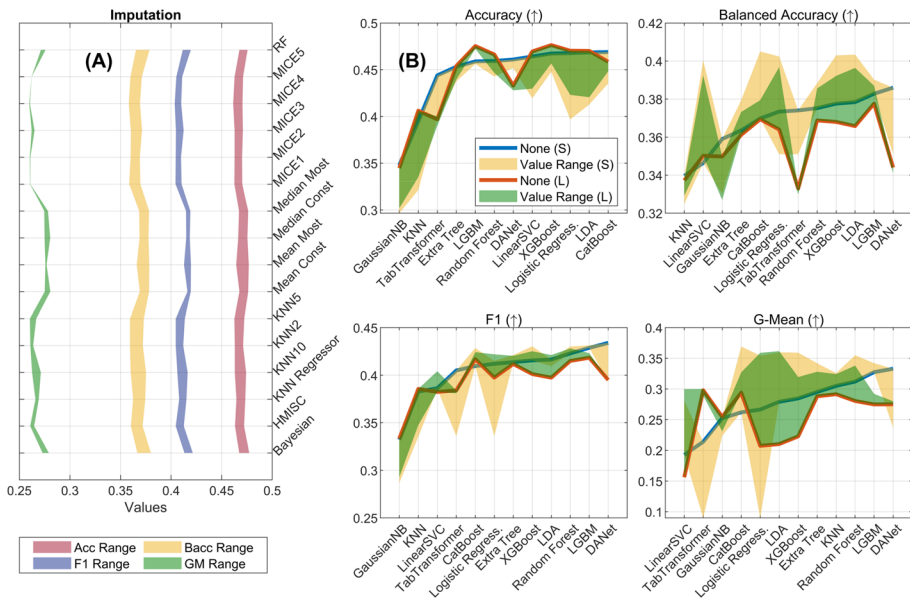


Fig. 12 Performance on the smaller (*S*) and larger (*L*) data for the 4-class prediction. **A** Effects of different data imputation techniques on various classifiers. **B** Performance metrics of various classifiers on the smaller dataset and the larger dataset using the full feature set and different data imbalance methods (sorted). *None* means no data imbalance. Best viewed in color

4.8 Summaries and Future Research

4.8.1 Are OSA Severity Predictions Hard?

With the largest OSA dataset in the literature and a wide range of classifiers, including SOTA DL ones, the best acquired Acc, Bacc, F1, and GM scores for the 4-class problem are 0.4726, 0.4062, 0.4336, and 0.4090, respectively, regardless of feature selections and data imbalance correction techniques (as shown in Figs. 3, 6, and 9).

Classifying 4 severity levels is clearly a difficult task as observed in previous works (Mencar et al., 2020; Rodrigues Jr et al., 2020; Ustun et al., 2016). Predicting binary classes is slightly easier with higher performances, e.g., the best acquired Bacc are 0.6948, 0.6780, and 0.6880 for AHI thresholds 5, 15, and 30, respectively. However, these are still weak numbers overall.

The complex relationships between OSA, demographics, symptoms, comorbidity and risk factors may provide an explanation for these classification challenges (Young et al., 2004; Romero-Corral et al., 2010; Peppard et al., 2000; Gottlieb & Punjabi, 2020; Li et al., 2023; Randerath et al., 2018). For example, while *diurnal somnolence* is among the most common symptoms, many OSA patients are asymptomatic (Gottlieb & Punjabi, 2020). For a deeper understanding, in Tables 6 and 7 in Appendices, we show the statistics for different severity classes and the full data. Though various trends are observed, the range values are highly overlapped among all 4 classes (refers to Sect. 4.1.1 Data Characteristics). These overlaps show the highly complicated nature of OSA that challenges classifiers to clarify class boundaries.

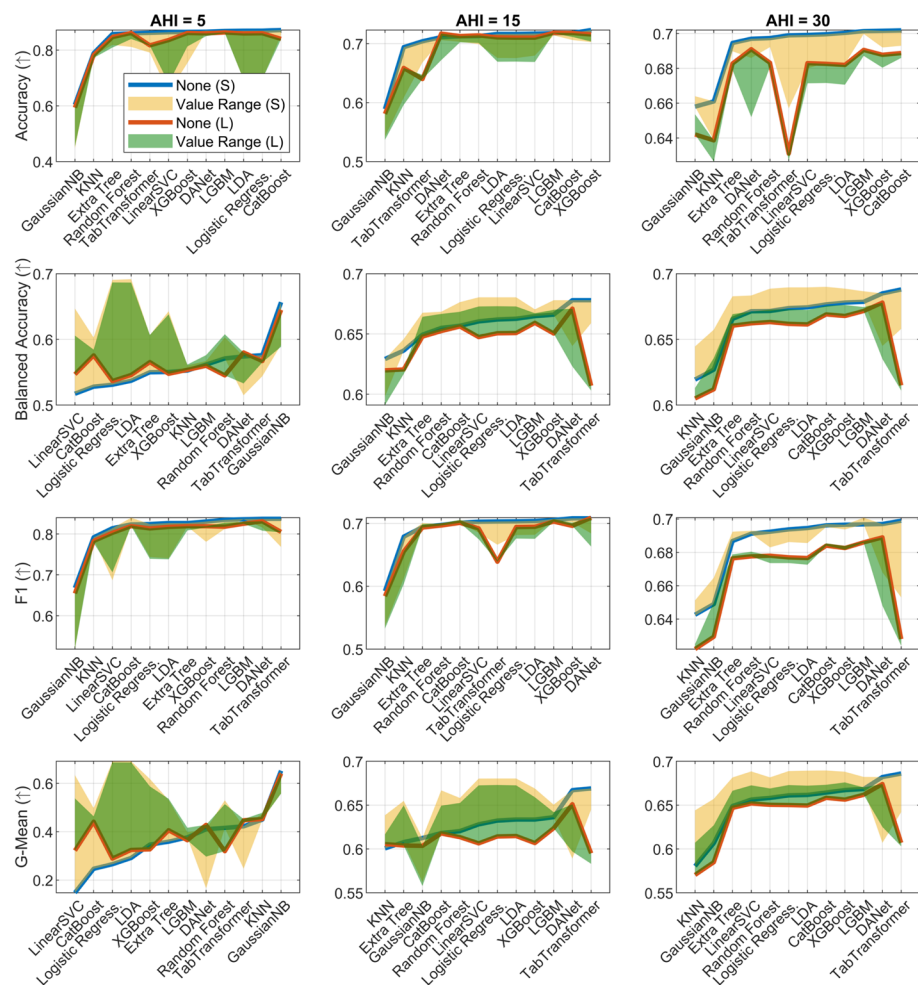


Fig. 13 Performances on the smaller (*S*) and larger (*L*) data for the binary predictions with $AHI=\{5, 15, 30\}$ using the full feature set and different data imbalance methods (sorted). *None* means no data imbalance. Best viewed in color

Moreover, AHI is used to define OSA severity levels in all existing works with common thresholds of 5, 15, and 30 (Mencar et al., 2020; Li et al., 2023; Huang et al., 2020a; Randerath et al., 2018). However, concerns have recently been raised regarding its limitations in assessing OSA (Randerath et al., 2018). In support of these concerns, directly predicting AHI values did not lead to better results in our analysis, and resulted in large errors (Sect. 4.7).

4.8.2 ML or DL?

Though DL methods tend to perform better than ML ones across all settings, there are not many differences in performance between them in our experiments. Among studied meth-

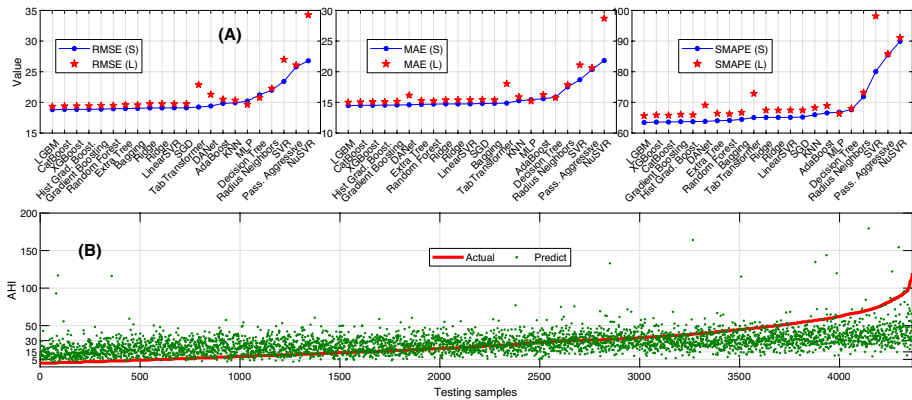


Fig. 14 **A** Performances of various regressors (RMSE, MAE and SMAPE) wrt. smaller (S) and larger (L) datasets. **B** Actual and predicted values on a test set for DANet (sorted by the actual values). Best viewed in color

ods, DANet and LGBM are top techniques for DL and ML, respectively, over all classification settings and evaluation metrics as demonstrated in Sect. 4.3. Moreover, taking into account the huge training times of DL, ML methods seems to be more tempting. However, DL approaches, with their proven unveiling powers in many challenging tasks (Alzubaidi et al., 2021), may be able to break the current performance barriers if large, complete, high-quality, and feature-rich datasets are combined with more sophisticated novel model designs and advanced learning technologies such as transfer learning (Zhuang et al., 2020; Van Der Aar et al., 2024) or knowledge distillation (Gou et al., 2021).

4.8.3 Imbalance Data Distributions

For all classification scenarios, clear biases toward *mild* and (especially) *severe* OSA classes are commonly seen as shown in Figs. 3 and 4. One reason may be the domination of high-severity samples in all existing datasets (Mencar et al., 2020; Rodrigues Jr et al., 2020). However, attempts to fix the problem by using synthetic data generation techniques like SMOTE or ADASYN (He et al., 2008) or by reducing data samples with undersampling (He et al., 2008) to balance the data distributions only help to improve the performance of various weak classifiers while having very slight effects on stronger ones as shown in Figs. 9 and 10. Hence, apart from balancing data distributions, attempts should also be made to explore and develop new classifiers tailored to deal with imbalanced input data.

4.8.4 More or Less Features?

While existing research usually employs different feature selection methods to enhance performance, our experiments on a large dataset with very diverse features show that they should be used with caution since most of the feature sets lead to significant performance drops over all models and metrics. Improvements are only observed on weak classifiers, while stronger classifiers seems to work better with the full feature set as shown in Figs. 6 and 7. Hence, while it does not mean we should not explore feature subsets since it is meaningful to have simpler

clinical screening tools for OSA such as Howarth et al. (2025), some efforts should also be spent on studying more diverse classifiers.

Moreover, while our OSFP dataset covers most common OSA related features known in the literature, they seem to be not enough due to very low prediction performances shown in Sect. 4. Hence, including more effective clinical features would be an important direction to improve the prediction accuracy, e.g., sleeping times (Huang et al., 2020a).

4.8.5 Multi-modal Data?

Since pure clinical data has limitations as studied above, combining them with other kinds of data (including PSG and non-PSG ones) into multi-modal classifiers would help to improve the assessment accuracy. For example, Lee et al. (2024) combine clinical features with polysomnography, X-ray and MRI data for enhancing prediction accuracy. Similarly, Qiu et al. (2026) propose the MIMAR-OSA framework that combines three modalities including basic physiological measurements, facial image and snoring audio to predict OSA with 92.24% accuracy. Another interesting direction would be using clinical features as a filter to identify patients with OSA before further clarifying their severity levels with some PSG-signals such as Kuo et al. (2024).

5 Conclusion

In this paper, we review existing machine learning approaches for predicting OSA severity. From the problem perspective, OSA severity prediction are commonly in the forms of severity level classifications (with 4 main severity classes or binary classes using different AHI thresholds) and AHI regression to predict AHI values directly. From the data perspective, existing works rely on three main kinds of data: PSG signals, Non-PSG signals and clinical data.

While using signals results in high accurate predictions, clinical data attract more attention due to their ease to collect. However, existing works report highly variation accuracy on datasets with small sizes and features. Hence, we conduct extensive studies with 39 different ML and DL classifiers/regressors combined together with 17 different sets of features, 4 data imbalance correction techniques, and 16 data imputation methods to re-evaluate existing results on a large scale, diverse and less-biased dataset with 20K to 80K samples and 51 commonly used features spanning through all relevant factors of OSA. Many of these techniques are commonly used by existing works while others, such as DL models like DANet or NODE, have not been studied before. The study reveals the complexity of OSA predictions, with models showing low prediction accuracy and biases towards severe OSA. While common techniques for boosting accuracy (e.g., feature selections or imbalance corrections) only have very small effects on performance, they may be able to lead to simpler, more accessible clinical tools. Looking forwards, larger and cleaner datasets with sophisticated DL models and learning strategies (e.g. development of multi-modal approaches combining signals and clinical data, or exploration of personalised models) may be one way to break current performance barriers.

5.1 Limitations and Future Works

Currently, our comparative study is only limited in EHR clinical data. Exploring performances of different approaches on large-scale PSG and other data is one of our future works.

Appendix A: Further Information

Data Imputation Methods. Simple imputer from scikit-learn was used to impute missing data using 4 methods: mean (continuous)/0 (categorical), mean (continuous)/mode (categorical), median (continuous)/0 (categorical), median (continuous)/mode (categorical). Categorical features had values of 0 or 1 for absence or presence of the condition, respectively, and 0 was chosen as the constant value replacement for missing values. The KNN imputer from scikit-learn fills missing values based on a user-specified number of nearest neighbors to a missing value. Three different values of k were 2, 5, and 10. The Iterative Imputer (also from scikit-learn) uses all features to predict missing values. The feature column with missing values is taken as output, y , and all other feature columns are taken as input, x . A regressor is then fit onto (x, y) for known y values and subsequently used to predict the missing values of y . Three different regressors were trialed: Bayesian Ridge, RF, and k -neighbors. Multivariate Imputation by Chained Equations (MICE) is obtained from R 4.1.1 with different numbers of iterations from 5 to 20 (Tables 6, 7, 8, 9, 10, 11 and 12).

Data Availability. The OSFP dataset can be accessed upon request to the investigator of the cohort and with a specific request regarding the French Law (MR004 document). Requests can be sent via the OSFP website (<https://www.osfp.fr/>).

Source Codes. Source codes of the paper can be found at: https://github.com/Duckie-Ducks/OSA_prediction.

Table 6 Patients characteristics of the entire cohort and by Severity (in real number: following with min, max, **mean** and **standard deviation**)

Feature Name	All	Normal	Mild	Moderate	Severe
		AHI < 5	5 ≤ AHI < 15	15 ≤ AHI < 30	AHI ≥ 30
	(N = 21818)	(N = 2862)	(N = 5036)	(N = 5142)	(N = 8778)
Demographics					
Age	18.03, 111.87, 53.26, 13.46	18.30, 86.46, 46.19, 13.24	18.03, 90.80, 50.56, 13.31	18.23, 91.21, 53.59, 13.07	18.88, 111.87, 56.92, 12.54
Height	116.00, 205.00, 169.70, 9.35	140.00, 200.00, 167.98, 8.94	137.00, 198.00, 168.88, 9.32	140.00, 205.00, 169.67, 9.54	116.00, 205.00, 170.76, 9.26
Weight	34.00, 200.00, 89.19, 19.89	40.00, 162.00, 82.04, 19.20	35.00, 188.00, 85.27, 18.84	34.00, 181.00, 87.58, 18.53	42.00, 200.00, 94.73, 20.00
Measurements					
Cervical perimeter	25.00, 60.00, 40.15, 4.56	25.00, 60.00, 37.88, 4.33	28.00, 60.00, 38.86, 4.23	26.00, 60.00, 39.65, 4.06	28.00, 60.00, 41.92, 4.44
Abdominal perimeter	60.00, 180.00, 106.05, 15.79	60.00, 157.00, 98.84, 15.84	60.00, 170.00, 102.20, 15.17	60.00, 167.00, 104.63, 14.52	61.00, 180.00, 111.43, 15.07
Systolic BP	0.00, 230.00, 132.49, 20.35	10.00, 202.00, 126.95, 17.98	0.00, 220.00, 129.90, 19.94	10.00, 223.00, 132.29, 20.42	10.00, 230.00, 135.91, 20.64
Diastolic BP	0.00, 181.00, 79.68, 13.74	6.00, 155.00, 76.66, 12.13	5.00, 158.00, 78.41, 13.46	6.00, 140.00, 79.78, 13.79	0.00, 181.00, 81.33, 14.13
BMI	12.64, 59.99, 30.99, 6.62	16.02, 58.30, 29.10, 6.68	14.57, 59.58, 29.95, 6.59	12.64, 59.17, 30.49, 6.40	16.00, 59.99, 32.49, 6.44
Questionnaires					
Epworth scale	0.00, 24.00, 9.64, 5.12	0.00, 24.00, 9.56, 5.14	0.00, 24.00, 9.27, 5.06	0.00, 24.00, 9.51, 5.09	0.00, 24.00, 9.95, 5.14
Pichots scale	0.00, 32.00, 13.39, 8.39	0.00, 32.00, 14.55, 8.45	0.00, 32.00, 13.70, 8.51	0.00, 32.00, 13.40, 8.36	0.00, 32.00, 12.82, 8.27
Depression scale	0.00, 13.00, 4.26, 3.84	0.00, 13.00, 4.60, 3.86	0.00, 13.00, 4.38, 3.90	0.00, 13.00, 4.25, 3.79	0.00, 13.00, 4.08, 3.81

Table 7 Patients characteristics of the entire cohort and by Severity (in binary: number (percentage %))

Feature Name	All	Normal	Mild	Moderate	Severe
		AHI < 5	5 ≤ AHI < 15	15 ≤ AHI < 30	30 ≤ AHI
	(N = 21818)	(N = 2862)	(N = 5036)	(N = 5142)	(N = 8778)
Demographic					
Male	13540(62.06)	1253(43.78)	2671(53.04)	3162(61.49)	6454(73.52)
Current smoker	3835(17.58)	547(19.11)	926(18.39)	880(17.11)	1482(16.88)
Former smoker	5849(26.81)	484(16.91)	1139(22.62)	1434(27.89)	2792(31.81)
Sedentary	2768(12.69)	307(10.73)	577(11.46)	573(11.14)	1311(14.94)
Measurements					
Maxillofacial profile	1529(7.01)	168(5.87)	339(6.73)	417(8.11)	605(6.89)
High BP	8293(38.01)	653(22.82)	1595(31.67)	1932(37.57)	4113(46.86)
Comorbidities					
Asthma	2533(11.61)	412(14.40)	653(12.97)	649(12.62)	819(9.33)
Rhinitis	2211(10.13)	356(12.44)	560(11.12)	520(10.11)	775(8.83)
COPD	891(4.08)	95(3.32)	171(3.40)	212(4.12)	413(4.70)
Respiratory fail	285(1.31)	25(0.87)	49(0.97)	70(1.36)	141(1.61)
Myocardial infarct	617(2.83)	38(1.33)	105(2.08)	128(2.49)	346(3.94)
Coronary fail	781(3.58)	42(1.47)	117(2.32)	190(3.70)	432(4.92)
Arrhythmias	1482(6.79)	120(4.19)	269(5.34)	365(7.10)	728(8.29)
Stroke	656(3.01)	47(1.64)	107(2.12)	170(3.31)	332(3.78)
Heart fail	408(1.87)	20(0.70)	55(1.09)	95(1.85)	238(2.71)
Arteriopathy	280(1.28)	25(0.87)	40(0.79)	73(1.42)	142(1.62)
Gastric reflux	3723(17.06)	394(13.77)	769(15.27)	974(18.94)	1586(18.07)
Glaucoma	437(2.00)	48(1.68)	110(2.18)	99(1.93)	180(2.05)
Diabetes	2721(12.47)	205(7.16)	461(9.15)	636(12.37)	1419(16.17)
Hypercholesterolemia	5166(23.68)	417(14.57)	972(19.30)	1259(24.48)	2518(28.69)
Hypertriglyceridemia	1010(4.63)	72(2.52)	168(3.34)	242(4.71)	528(6.02)
Hypo(er)thyroidism	1673(7.67)	224(7.83)	389(7.72)	434(8.44)	626(7.13)
Depression	2942(13.48)	366(12.79)	695(13.80)	698(13.57)	1183(13.48)
Obesity	4909(22.50)	723(25.26)	1181(23.45)	1082(21.04)	1923(21.91)
Dysmorphology	137(0.63)	12(0.42)	33(0.66)	30(0.58)	62(0.71)
Restless Leg Syndrome	3758(17.22)	490(17.12)	894(17.75)	901(17.52)	1473(16.78)
Symptoms					
Snoring	21120(96.80)	2739(95.70)	4835(96.01)	4957(96.40)	8589(97.85)
Diurnal somnolence	17841(81.77)	2329(81.38)	4031(80.04)	4205(81.78)	7276(82.89)
Driving drowsiness	9862(45.20)	1394(48.71)	2340(46.47)	2329(45.29)	3799(43.28)
Morning fatigue	16734(76.70)	2275(79.49)	3917(77.78)	3972(77.25)	6570(74.85)
Morning headache	10290(47.16)	1578(55.14)	2543(50.50)	2505(48.72)	3664(41.74)
Memory problem	3213(14.73)	307(10.73)	619(12.29)	758(14.74)	1529(17.42)
Nocturnal perspiration	11904(54.56)	1677(58.60)	2853(56.65)	2851(55.45)	4523(51.53)
Shortness of breath on exertion	15064(69.04)	2014(70.37)	3496(69.42)	3559(69.21)	5995(68.30)
Nocturia	12470(57.15)	1368(47.80)	2582(51.27)	2895(56.30)	5625(64.08)
Drowsiness accident	540(2.48)	55(1.92)	118(2.34)	135(2.63)	232(2.64)
Near miss accident	1953(8.95)	276(9.64)	447(8.88)	438(8.52)	792(9.02)
Respiratory arrest	9659(44.27)	870(30.40)	1748(34.71)	2184(42.47)	4857(55.33)

Table 8 Features belonging to each feature-selection group

Features	Literature					Category				Statistics							Cat	SR	
	All	Wu	M.	Hu.	Us.	Ro.	De.	Me.	Co.	Sy.	Qu.	Kb	Kb	RI	RP	Cor			Kr
Sex	✓		✓		✓		✓					✓	✓				✓	✓	✓
Age	✓	✓		✓	✓	✓	✓						✓	✓	✓	✓	✓	✓	✓
Current smoker	✓			✓	✓	✓	✓												
Former smoker	✓			✓	✓	✓	✓					✓							✓
Sedentary	✓						✓												
Height	✓							✓						✓					✓
Weight	✓							✓					✓	✓	✓	✓	✓		✓
Cervical perimeter	✓	✓		✓		✓		✓				✓	✓	✓	✓	✓	✓	✓	✓
Abdominal perimeter	✓	✓			✓	✓		✓				✓	✓	✓	✓	✓	✓	✓	✓
Systolic BP	✓	✓						✓					✓	✓		✓		✓	✓
Diastolic BP	✓	✓				✓		✓					✓	✓	✓		✓	✓	✓
Maxillofacial profile	✓							✓							✓				
BMI	✓	✓	✓		✓	✓		✓					✓	✓	✓	✓	✓	✓	✓
High BP	✓				✓			✓								✓			
Asthma	✓								✓				✓			✓	✓		
Rhinitis	✓								✓										
COPD	✓								✓										
Respiratory fail	✓								✓			✓							
Myocardial infarct	✓								✓										
Coronary fail	✓								✓			✓							
Arrhythmias	✓								✓										
Stroke	✓								✓										
Heart Fail	✓								✓										
Arteriopathy	✓								✓										
Gastric reflux	✓								✓										
Glaucoma	✓								✓										
Diabetes	✓			✓					✓			✓			✓				

Table 8 (continued)

Features	Literature						Category			Statistics									
	All	Wu	M.	Hu.	Us.	Ro.	De.	Me.	Co.	Sy.	Qu.	Kb	Kb	RI	RP	Cor	Kr	Cat	SR
Hypercholesterolemia	✓								✓			✓				✓			
Hypertriglyceridemia	✓								✓										
Hypo(er)thyroidism	✓								✓										
Depression	✓						✓		✓										
Obesity	✓								✓										
Dysmorphology	✓								✓										
Restless Leg Syndrome	✓								✓										
Snoring	✓			✓						✓									
Diurnal somnolence	✓									✓									
Driving drowsiness	✓									✓									
Morning fatigue	✓									✓									
Morning headache	✓									✓		✓							
Memory problem	✓									✓									
Nocturnal perspiration	✓									✓									
Shortness of breath on exertion	✓									✓									
Nocturia	✓					✓						✓				✓			
Drowsiness accident	✓									✓									
Near Miss accident	✓									✓									
Respiratory arrest	✓									✓		✓			✓	✓	✓	✓	
Epworth scale	✓	✓									✓								
Pichots scale	✓										✓			✓					
Depression scale	✓					✓					✓			✓					✓

M.: Mencar, Hu.: Huang, Us.: Ustun, Ro.: Rodruiges, De.: Demographic, Me.: Measurements, Co.: Comorbidities, Sy.: Symptoms, Qu.: Questionnaires, Cor: Correlation, RI: RF Impurity, RP: RF Permutation, Kb: Kbest Fclass, Kb: Kbest Chi Square, Kr: Kruskal Chi Square, Cat: CatBoost, SR: SHAP RF

Table 9 Description of features in the dataset

Features	Type	Description
Sex	Categorical	Patient's sex: Male(1)/Female(2)
Age	Continuous	Patient's age (unit: years)
Current smoker	Binary	Being a current smoker: Yes(1)/No(0)
Former smoker	Binary	Being a former smoker: Yes(1)/No(0)
Sedentary	Binary	Being physically inactive: Yes(1)/No(0)
Height	Continuous	Patient's height (unit: cm)
Weight	Continuous	Patient's weight (unit: cm)
Cervical perimeter	Continuous	Circumference around the neck region
Abdominal perimeter	Continuous	Circumference around the abdomen
Systolic BP	Continuous	Maximum arterial pressure during heart contraction
Diastolic BP	Continuous	Arterial pressure during heart relaxation
Maxillofacial profile	Binary	Presence of maxillofacial profile alteration: Yes(1)/No(0)
BMI	Continuous	Body mass index (BMI)
High BP	Binary	High blood pressure: Yes(1)/No(0)
Rhinitis	Binary	History of Rhinitis: Yes(1)/No(0)
COPD	Binary	History of COPD: Yes(1)/No(0)
Respiratory fail	Binary	History of respiratory failure: Yes(1)/No(0)
Myocardial infarct	Binary	History of myocardial infarction: Yes(1)/No(0)
Coronary fail	Binary	History of coronary failure: Yes(1)/No(0)
Arrhythmias	Binary	History of Arrhythmias: Yes(1)/No(0)
Stroke	Binary	History of stroke: Yes(1)/No(0)
Heart Fail	Binary	History of heart failure: Yes(1)/No(0)
Arteriopathy	Binary	History of arteriopathy: Yes(1)/No(0)
Gastric reflux	Binary	History of gastric reflux: Yes(1)/No(0)
Glaucoma	Binary	History of glaucoma: Yes(1)/No(0)
Diabetes	Binary	History of diabetes: Yes(1)/No(0)
Hypercholesterolemia	Binary	History of hypercholesterolemia: Yes(1)/No(0)
Hypertriglyceridemia	Binary	History of hypertriglyceridemia: Yes(1)/No(0)
Hypo(er)thyroidism	Binary	History of thyroid disorder: Yes(1)/No(0)
Depression	Binary	History of depression: Yes(1)/No(0)
Obesity	Binary	History of obesity: Yes(1)/No(0)
Dysmorphology	Binary	Presence of dysmorphic features: Yes(1)/No(0)
Restless Leg Syndrome	Categorical	Restless Legs Syndrome diagnostic criteria
Snoring	Categorical	Snoring scale
Diurnal somnolence	Binary	Presence of diurnal somnolence: Yes(1)/No(0)
Driving drowsiness	Binary	Experience of drowsiness while driving: Yes(1)/No(0)
Morning fatigue	Binary	Presence of morning fatigue: Yes(1)/No(0)
Morning headache	Binary	Presence of morning headache: Yes(1)/No(0)
Memory problem	Binary	Presence of memory problem: Yes(1)/No(0)
Nocturnal perspiration	Binary	Presence of nocturnal perspiration: Yes(1)/No(0)
Shortness of breath on exertion	Binary	Experience of shortness of breath on exertion: Yes(1)/No(0)
Nocturia	Binary	Presence of nocturia: Yes(1)/No(0)
Drowsiness accident	Binary	History of drowsiness-related accident: Yes(1)/No(0)
Near Miss accident	Binary	History of near-miss accident: Yes(1)/No(0)
Respiratory arrest	Binary	History of respiratory arrest: Yes(1)/No(0)
Epworth scale	Categorical	Epworth Sleepiness Scale score
Pichots scale	Categorical	Pichot scale score
Depression scale	Categorical	Depression scale score
Apnea hypopnea index	Continuous	Apnea-Hypopnea Index (events/hour)

Table 10 Search space for hyperparameter tuning of 12 deep learning models

Model	Hyperparameters
DANet	max_epochs = [50, 100, 150] learning_rate = [1e-4, 5e-4, 1e-3, 1e-2, 5e-2]
TabTransformer	max_epochs = [150] learning_rate = [1e-3, 5e-3, 1e-2]
LSTM	hidden_dim = [50, 100, 200] learning_rate = [0.01, 0.05, 0.1, 0.5] n_layers = [1, 2, 3]
GATE	max_epochs = [50, 100, 150] learning_rate = [1e-4, 5e-4, 1e-3, 5e-3, 1e-2, 5e-2]
CEM	max_epochs = [50, 100, 150] learning_rate = [1e-3, 5e-3, 1e-2]
NODE	max_epochs = [50, 100, 150] learning_rate = [1e-4, 5e-4, 1e-3, 5e-3, 1e-2, 5e-2]
RNN	hidden_dim = [50, 100, 200] learning_rate = [0.01, 0.05, 0.1, 0.5] n_layers = [1, 2, 3]
Deep Belief Network	hidden_layers_structure = [(50,30,10),(50,30),50,30] learning_rate_rbm = [0.001, 0.01, 0.1] learning_rate = [0.001, 0.01, 0.1] n_epochs_rbm = [10, 15, 20] batch_size = [32, 64, 128] activation_fn = ['relu', 'sigmoid'] dropout_p = [0.2, 0.5]
TabNet	max_epochs = [50, 100, 150] learning_rate = [1e-3, 5e-3, 1e-2]
CNN	dropout = [0.01, 0.05, 0.1, 0.2, 0.5] learning_rate = [0.001, 0.01, 0.05, 0.1, 0.5]
GCN	hidden_dim = [8, 16] n_epochs = [20, 50] n_layers = [1, 2, 3] learning_rate = [0.0001, 0.01]
GRU	hidden_dim = [20, 50] n_layers = [1, 2] learning_rate = [0.0001, 0.01]

Table 11 Search space for hyperparameter tuning of 27 Machine Learning models

Model	Hyperparameters
Nearest Centroid	metric = ['euclidean', 'manhattan', 'l1', 'l2', 'nan_euclidean', 'cosine']
Hist Gradient Boosting	learning_rate = [0.0001, 0.001, 0.01, 0.1, 1] max_leaf_nodes = [5, 10, 30]
Gradient Boosting	n_estimators = [10, 50, 100] learning_rate = [0.00001, 0.0001, 0.001, 0.01, 0.1, 1]
LGBM	num_iterations = [150, 200, 500] learning_rate = [0.05, 0.005] max_depth = [10, 15, 25] num_leaves = [7, 15, 31] bagging_freq = [100, 200, 400]
MLP	hidden_layer_sizes = [[50, 30, 10], [50, 30], 50, 30, 10] activation = ['tanh', 'relu'] alpha = [0.0001, 0.001, 0.01, 0.1] solver = ['sgd', 'adam'] learning_rate = ['adaptive', 'constant', 'invscaling']
Bagging	n_estimators = [10, 50, 100, 500] max_features = [0.6, 0.7, 0.8, 0.9, 1.0] bootstrap = [true, false] oob_score = [true, false]
Random Forest	max_depth = [10, 30, 50, 70, 90, null] n_estimators = [10, 50, 100] max_features = ['auto', 'sqrt']
SGD	loss = ['hinge', 'log', 'modified_huber', 'squared_hinge', 'perceptron'] penalty = ['l1', 'l2', 'elasticnet', 'none'] alpha = [1e-5, 1e-4, 1e-3, 1e-2, 1e-1]
XGBoost	n_estimators = [10, 50, 100] learning_rate = [0.001, 0.01, 0.05] reg_alpha = [0.5, 1]
GaussianNB	var_smoothing = [0.01, 0.001, 0.0001, 1e-6, 1e-8, 1e-12, 1e-15]
AdaBoost	n_estimators = [10, 50, 100, 500] learning_rate = [0.1, 1.0, 2.0, 10.0]
LDA	solver = ['lsqr', 'eigen'] n_components = [1, 5, 10, 50]
MultinomialNB	alpha = [0, 0.3, 0.5, 1.0] force_alpha = [true, false] fit_prior = [true, false]
Logistic Regression	penalty = ['l1', 'l2', 'elasticnet', 'none'] C = [1, 10, 100, 1000] solver = ['lbfgs', 'newton-cg', 'liblinear', 'sag', 'saga']
CatBoost	depth = [4, 5, 6, 7, 8, 9, 10] learning_rate = [0.01, 0.02, 0.03, 0.04] iterations = [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]
Extra Trees	n_estimators = [10, 50, 100, 500] max_depth = [10, 30, 50, 70, 90, null] min_samples_split = [2, 5, 10] min_samples_leaf = [1, 2, 4]
BernoulliNB	alpha = [0, 0.3, 0.5, 1] force_alpha = [true, false] fit_prior = [true, false]
QDA	reg_param = [0.1, 0.2, 0.3, 0.4, 0.5]
LinearSVC	C = [1, 10, 100, 1000] penalty = ['l1', 'l2', 'elasticnet', 'none']
Decision Tree	max_depth = [10, 30, 50, 70, 90, null] min_sample_leaf = [1, 2, 4] criterion = ['gini', 'entropy', 'log_loss']
Ridge	alpha = [0.1, 1, 2, 5, 10] fit_intercept = [true, false]
KNN	n_neighbors = [1, 10, 100] leaf_size = [1, 20, 40, 100] p = [1, 2]
Perceptron	penalty = ['l1', 'l2', 'elasticnet', 'none'] alpha = [1e-5, 1e-4, 1e-3, 1e-2, 1e-1]
Passive Aggressive	C = [1, 10, 100, 1000] fit_intercept = [true, false]
NuSVC	gamma = [0.001, 0.0001] kernel = ['linear', 'poly', 'rbf', 'sigmoid'] nu = [0.1, 0.2, 0.3, 0.4]
SVC	C = [1, 10, 100, 1000] kernel = ['linear', 'rbf'] gamma = [0.001, 0.0001]
Radius Neighbors	weights = ['uniform', 'distance'] algorithm = ['auto', 'ball_tree', 'kd_tree', 'brute'] leaf_size = [20, 40, 100, 1]

Table 12 Means and standard deviations of performance metrics for all methods using all features

	Mean				Standard deviation			
	Acc	F1	Bacc	GM	Acc	F1	Bacc	GM
Nearest Centroid	0.3812	0.3194	0.3638	0.2859018	0.003347	0.007537	0.009445	0.011515
Logistic Regression	0.4678	0.354	0.3734	0.2664634	0.004658	0.007778	0.005941	0.010573
SGD	0.375	0.3318	0.338	0.3142854	0.014248	0.011713	0.009539	0.009433
Perceptron	0.3698	0.265	0.3282	0.1442898	0.081846	0.03532	0.033544	0.113469
Decision Tree	0.4272	0.353	0.3568	0.3082582	0.005762	0.003082	0.004147	0.00724
Bagging	0.4494	0.3686	0.3726	0.3159492	0.003912	0.005899	0.005683	0.00693
Extra Trees	0.4532	0.3566	0.3636	0.295167	0.007463	0.006465	0.005941	0.008732
Hist Gradient Boosting	0.4588	0.3688	0.3776	0.308169	0.0111	0.005762	0.007797	0.005888
LDA	0.4686	0.3614	0.3784	0.2791952	0.005505	0.007503	0.006269	0.008543
QDA	0.4232	0.3376	0.3742	0.2529086	0.004025	0.004278	0.006301	0.013025
BernoulliNB	0.3774	0.311	0.355	0.2694734	0.028148	0.009721	0.010173	0.025414
MultinomialNB	0.3646	0.305	0.3578	0.2714834	0.002302	0.006364	0.006419	0.012232
GaussianNB	0.3478	0.2934	0.359	0.2522892	0.009257	0.015209	0.009192	0.022852
MLP	0.46	0.3694	0.375	0.310049	0.008631	0.007021	0.003742	0.013858
SVC	0.2648	0.243	0.256	0.2419614	0.024692	0.004899	0.001225	0.01618
LinearSVC	0.4642	0.309	0.3464	0.1924998	0.003114	0.006481	0.000548	0.015615
NuSVC	0.1914	0.158	0.266	0.0143668	0.036569	0.035644	0.033867	0.032125
XGBoost	0.4674	0.362	0.3776	0.2847226	0.005079	0.007211	0.007503	0.004804
Random Forest	0.4594	0.3702	0.3752	0.312086	0.008503	0.009365	0.007887	0.010086
AdaBoost	0.4636	0.3686	0.3792	0.3044946	0.003782	0.004827	0.003271	0.006523
Gradient Boosting	0.4578	0.3684	0.376	0.308312	0.00614	0.008019	0.008337	0.006814
Radius Neighbors	0.402	0.143	0.25	0	0	0	0	0
Ridge	0.4634	0.2846	0.3394	0.1312368	0.002881	0.001949	0.001517	0.009557
Passive Aggressive	0.3462	0.282	0.323	0.2299	0.081048	0.05652	0.017649	0.077021
LGBM	0.459	0.3788	0.3828	0.327516	0.005958	0.008871	0.008349	0.010008
CatBoost	0.469	0.3516	0.3698	0.2620074	0.005148	0.008234	0.006648	0.009355
KNN	0.395	0.3362	0.3396	0.3051086	0.006245	0.00507	0.005941	0.005586
CEM	0.4626	0.3774	0.3838	0.3172	0.004879	0.005273	0.002683	0.014772
DANet	0.4614	0.3838	0.3864	0.3334	0.009072	0.004147	0.005595	0.00503
Node	0.4726	0.3586	0.3798	0.2646	0.007893	0.011059	0.006648	0.018942
GATE	0.4658	0.3782	0.3932	0.3084	0.011692	0.014721	0.006535	0.043592
TabNet	0.4346	0.2756	0.316	0.1582	0.015043	0.023223	0.016956	0.034687
TabTransformer	0.444	0.3462	0.3744	0.2144	0.011895	0.022521	0.010877	0.093385
LSTM	0.4484	0.37	0.3772	0.3202	0.010761	0.010173	0.00795	0.023317
GCN	0.4286	0.2568	0.3066	0.1162	0.00695	0.025273	0.016319	0.070201
RNN	0.451	0.3454	0.3722	0.2588	0.008972	0.016891	0.010569	0.042552
Deep Belief Network	0.4044	0.297	0.3292	0.2032	0.015884	0.014967	0.016453	0.035682
GRU	0.4588	0.3672	0.3832	0.2952	0.006458	0.018336	0.005891	0.053392
CNN	0.4028	0.147	0.251	0	0.001304	0.006164	0.002236	0

Additional Results on The SHHS-1 Dataset. Tables 13, 14, 15, 16 show the prediction accuracy of different methods on the SHHS-1 dataset for 4-classes and binary prediction tasks using 27 selected features as indicated in Table 17. Compared to the OSFP data, the SHHS-1 dataset is much smaller with only 4,218 patients and is thus less generic. Though there are deviations in detailed performances due to data differences, but the messages are

Table 13 Means and standard deviations of performance metric values for all methods using the SHHS-1 dataset as an all classes prediction problem using 27 clinical features selected from the SHHS-1 cohort

	Mean				Standard deviation			
	Acc	F1	Bacc	GM	Acc	F1	Bacc	GM
Nearest Centroid	0.4392	0.4184	0.3882	0.315203	0.015561	0.017374	0.024087	0.026649
Logistic Regression	0.507	0.4554	0.321	0.178402	0.009529	0.005004	0.003225	0.015158
SGD	0.4458	0.4298	0.3322	0.280345	0.010245	0.011053	0.009704	0.022209
Perceptron	0.4372	0.4038	0.3282	0.104101	0.047474	0.036168	0.02691	0.095047
Decision Tree	0.442	0.43	0.3124	0.241275	0.014967	0.010257	0.0092	0.009759
Bagging	0.4706	0.4384	0.3138	0.213414	0.013396	0.012355	0.011071	0.02101
Extra Trees	0.4846	0.4398	0.3084	0.143073	0.012159	0.013303	0.012532	0.077279
Hist Gradient Boosting	0.4954	0.4584	0.3296	0.221477	0.014305	0.012043	0.009243	0.015688
LDA	0.509	0.4604	0.3268	0.186881	0.012313	0.008686	0.008818	0.022643
QDA	0.487	0.3822	0.274	0.044667	0.009143	0.013318	0.01064	0.055645
BernoulliNB	0.381	0.3766	0.347	0.28048	0.0229	0.024393	0.029967	0.038395
Multi-nomialNB	0.3842	0.3714	0.3604	0.264346	0.015263	0.014122	0.017749	0.021462
GaussianNB	0.4744	0.4444	0.359	0.297564	0.01693	0.020016	0.017029	0.024191
MLP	0.496	0.4628	0.3298	0.217359	0.022521	0.018936	0.016191	0.026544
SVC	0.3524	0.3622	0.277	0.237123	0.057608	0.050866	0.033365	0.030472
LinearSVC	0.499	0.4362	0.3018	0.027369	0.013682	0.014905	0.011053	0.054738
NuSVC	0.4694	0.411	0.3092	0.062369	0.026748	0.029833	0.026301	0.07677
XG-Boost	0.4954	0.4568	0.3252	0.205802	0.012579	0.014274	0.013378	0.02797
Random Forest	0.495	0.444	0.308	0.120838	0.010257	0.011713	0.011454	0.066365
Ada-Boost	0.4774	0.4454	0.3202	0.225014	0.01398	0.014651	0.016952	0.031913
Gradient Boosting	0.4862	0.4586	0.332	0.233867	0.009887	0.00403	0.012759	0.037323
Radius Neighbors	0.477	0.3086	0.2502	0	0.000632	0.00102	0.0004	0
Ridge	0.5028	0.4274	0.2928	0	0.008998	0.006086	0.004833	0
Passive Aggressive	0.4336	0.3528	0.3068	0.077887	0.042566	0.028562	0.020798	0.097479
LGBM	0.487	0.4514	0.3226	0.213955	0.008989	0.009351	0.009156	0.022122

Table 13 (continued)

	Mean				Standard deviation			
	Acc	F1	Bacc	GM	Acc	F1	Bacc	GM
Cat-Boost	0.4836	0.4518	0.3252	0.2224562	0.011976644	0.007520638	0.008840814	0.026011065
KNN	0.4738	0.4356	0.305	0.140631	0.019508	0.01662	0.01205	0.072316
CEM	0.4926	0.4522	0.3282	0.2058	0.017408	0.026453	0.010591	0.038494
DANet	0.4934	0.4622	0.331	0.189	0.007499	0.011686	0.011781	0.040363
Node	0.5072	0.454	0.3162	0.1216	0.014358	0.017018	0.01087	0.021068
GATE	0.4998	0.4558	0.3228	0.0444	0.016265	0.019651	0.013848	0.0888
TabNet	0.4926	0.4028	0.2832	0	0.013603	0.019934	0.011016	0
Tab-Transformer	0.4908	0.4268	0.3138	0.128	0.013318	0.033737	0.013761	0.065308
LSTM	0.4956	0.4232	0.2976	0.078286	0.011943	0.035504	0.024776	0.095881
GCN	0.4996	0.429	0.306	0.113671	0.015239	0.041881	0.028657	0.094348
RNN	0.4852	0.3588	0.2684	0	0.013467	0.056276	0.021209	0
Deep Belief Network	0.4898	0.384	0.2714	0	0.006911	0.006481	0.003929	0
GRU	0.4898	0.4548	0.3218	0.157529	0.0156	0.008773	0.013273	0.092659
CNN	0.4948	0.4336	0.3008	0.051878	0.01264	0.020165	0.014204	0.064775

Table 14 Means and standard deviations of performance metric values for all methods using the SHHS-1 dataset as a binary classes prediction problem (AHI=5 threshold) using 27 clinical features selected from the SHHS-1 cohort

	Mean				Standard deviation			
	Acc	F1	Bacc	GM	Acc	F1	Bacc	GM
Nearest Centroid	0.6466	0.6444	0.6496	0.644768	0.017072	0.018467	0.016032	0.019277
Logistic Regression	0.681	0.6804	0.6794	0.677535	0.009654	0.009222	0.009047	0.008493
SGD	0.6802	0.6774	0.6764	0.669953	0.007494	0.008452	0.008089	0.010697
Perceptron	0.556	0.4802	0.5638	0.402229	0.029475	0.069344	0.034695	0.131243
Decision Tree	0.626	0.6238	0.6228	0.619105	0.008989	0.008863	0.008863	0.009242
Bagging	0.6728	0.6718	0.6706	0.667948	0.014552	0.013877	0.014066	0.013391
Extra Trees	0.6686	0.6676	0.6662	0.663977	0.00602	0.005463	0.005344	0.004775
Hist Gradient Boosting	0.68	0.677	0.6764	0.670331	0.013387	0.01352	0.01372	0.014977
LDA	0.6792	0.6786	0.6774	0.675692	0.012303	0.011926	0.011741	0.011021
QDA	0.6522	0.652	0.654	0.652536	0.008795	0.009099	0.008764	0.008694
BernoulliNB	0.626	0.6256	0.6246	0.623646	0.01911	0.018521	0.018832	0.018522
Multi-nominalNB	0.632	0.6312	0.6302	0.629107	0.013624	0.013732	0.013732	0.013441
GaussianNB	0.6456	0.6424	0.65	0.642367	0.011289	0.012784	0.010918	0.013388
MLP	0.6748	0.6734	0.672	0.668915	0.010572	0.010594	0.010469	0.010671
SVC	0.4862	0.4832	0.4854	0.478784	0.038969	0.040926	0.038619	0.043556
LinearSVC	0.6804	0.6796	0.6782	0.676639	0.010632	0.010307	0.009867	0.009305
NuSVC	0.5812	0.5338	0.5724	0.463276	0.079658	0.126752	0.087276	0.205371
XGBoost	0.6692	0.6678	0.6664	0.663195	0.009621	0.010068	0.009728	0.010671
Random Forest	0.6668	0.666	0.6644	0.662265	0.005741	0.006229	0.006468	0.00724
AdaBoost	0.6792	0.6788	0.678	0.677132	0.012254	0.012123	0.011781	0.011311
Gradient Boosting	0.6778	0.6754	0.6742	0.669189	0.006493	0.005535	0.005492	0.003359
Radius Neighbors	0.575	0.5024	0.5576	0.411553	0.009317	0.008616	0.009091	0.008283
Ridge	0.6786	0.6778	0.6766	0.674604	0.010837	0.010107	0.009831	0.009168
Passive Aggressive	0.6062	0.567	0.6104	0.531494	0.055733	0.087925	0.046992	0.111638

Table 14 (continued)

	Mean				Standard deviation			
	Acc	F1	Bacc	GM	Acc	F1	Bacc	GM
LGBM	0.667	0.6644	0.6634	0.658946	0.011082	0.010929	0.010929	0.010606
Cat-Boost	0.6768	0.676	0.6746	0.6724812	0.009516302	0.009055385	0.008616264	0.007906368
KNN	0.6572	0.657	0.6562	0.655537	0.006112	0.006099	0.00581	0.005712
CEM	0.655	0.6548	0.654	0.6528	0.018407	0.018038	0.017788	0.017058
DANet	0.6608	0.6594	0.6598	0.6564	0.010572	0.010707	0.010419	0.011792
Node	0.6828	0.6812	0.6802	0.6768	0.007859	0.007626	0.006794	0.006882
GATE	0.6688	0.6646	0.666	0.6572	0.012319	0.009415	0.01	0.005776
TabNet	0.6318	0.6314	0.631	0.6294	0.035079	0.035842	0.036442	0.037945
Tab-Trans-former	0.6854	0.6836	0.6822	0.678	0.010268	0.009478	0.009261	0.00839
LSTM	0.644	0.6292	0.6378	0.60805	0.031642	0.050657	0.035723	0.076447
GCN	0.6574	0.654	0.6536	0.646405	0.011482	0.012474	0.012126	0.015601
RNN	0.5942	0.5862	0.5968	0.580588	0.051425	0.066028	0.047059	0.076203
Deep Belief Network	0.6364	0.6352	0.6338	0.631541	0.017511	0.016869	0.016975	0.016203
GRU	0.6518	0.6488	0.6506	0.643922	0.025662	0.0256	0.022123	0.022981
CNN	0.664	0.6622	0.6616	0.657762	0.011261	0.011161	0.01191	0.013396

Table 15 Means and standard deviations of performance metric values for all methods using the SHHS-1 dataset as a binary classes prediction problem (AHI=15 threshold) using 27 clinical features selected from the SHHS-1 cohort

	Mean				Standard deviation			
	Acc	F1	Bacc	GM	Acc	F1	Bacc	GM
Nearest Centroid	0.6802	0.7048	0.6572	0.655179	0.008635	0.007467	0.019177	0.019976
Logistic Regression	0.789	0.7394	0.558	0.382148	0.006164	0.006859	0.007616	0.015532
SGD	0.7076	0.7146	0.5954	0.560908	0.009351	0.004923	0.006591	0.014679
Perceptron	0.7822	0.719	0.5404	0.258835	0.00337	0.028475	0.042354	0.199808
Decision Tree	0.7314	0.7134	0.5508	0.448604	0.017716	0.00979	0.011053	0.031511
Bagging	0.7472	0.7224	0.5592	0.442048	0.024774	0.0092	0.016546	0.064435
Extra Trees	0.7752	0.7258	0.5434	0.358951	0.006046	0.007386	0.009372	0.025862
Hist Gradient Boosting	0.757	0.7236	0.5506	0.409438	0.017216	0.008309	0.01136	0.04307
LDA	0.7872	0.7358	0.5534	0.369385	0.004707	0.006013	0.007283	0.018536
QDA	0.7816	0.699	0.5084	0.16622	0.003007	0.003633	0.003555	0.024435
BernoulliNB	0.5978	0.6334	0.619	0.616952	0.007305	0.006621	0.023858	0.02091
MultinomialNB	0.5946	0.6302	0.6302	0.62651	0.004454	0.00402	0.018755	0.015409
GaussianNB	0.7372	0.741	0.6266	0.595271	0.007547	0.008319	0.014664	0.019937
MLP	0.7744	0.7456	0.5794	0.466686	0.015409	0.011038	0.009222	0.006152
SVC	0.4828	0.5188	0.4712	0.456549	0.102484	0.098211	0.043393	0.040763
LinearSVC	0.7884	0.7172	0.5298	0.2511	0.00403	0.016738	0.019031	0.100919
NuSVC	0.6482	0.6398	0.598	0.506147	0.150736	0.137906	0.041785	0.115237
XGBoost	0.777	0.73	0.549	0.374605	0.006099	0.007266	0.009055	0.022446
Random Forest	0.7874	0.732	0.5478	0.34962	0.004673	0.005692	0.007026	0.016136
AdaBoost	0.7746	0.739	0.5662	0.429574	0.006184	0.009675	0.014091	0.031767
Gradient Boosting	0.7452	0.726	0.5644	0.465553	0.009642	0.004472	0.0028	0.014905
Radius Neighbors	0.784	0.6896	0.5	0	0	0.00049	0	0
Ridge	0.785	0.7118	0.5228	0.241813	0.004147	0.009683	0.010722	0.0455
Passive Aggressive	0.7738	0.7116	0.5368	0.260006	0.020429	0.015794	0.043259	0.190297
LGBM	0.7726	0.7318	0.5544	0.398737	0.005122	0.007414	0.009912	0.027388
CatBoost	0.7506	0.73	0.5684	0.4694354	0.007445804	0.0069857	0.010229369	0.018084498
KNN	0.6906	0.6912	0.5438	0.47808	0.006216	0.006177	0.011565	0.021958
CEM	0.773	0.7328	0.5586	0.4044	0.005762	0.014497	0.022633	0.066692
DANet	0.7788	0.7346	0.5574	0.3922	0.007652	0.015028	0.019916	0.064297
Node	0.7916	0.7326	0.5484	0.339	0.011943	0.020451	0.0135	0.045109
GATE	0.786	0.7282	0.5496	0.3356	0.012264	0.026202	0.032004	0.109681
TabNet	0.7788	0.699	0.5084	0.1786	0.007194	0.009338	0.006468	0.019335
TabTransformer	0.7864	0.7396	0.565	0.3962	0.009708	0.024254	0.031987	0.089399

Table 15 (continued)

	Mean				Standard deviation			
	Acc	F1	Bacc	GM	Acc	F1	Bacc	GM
LSTM	0.7856	0.701	0.5124	0.107658	0.002059	0.016649	0.018249	0.139865
GCN	0.783	0.7156	0.539	0.246183	0.007563	0.029042	0.051718	0.198741
RNN	0.784	0.6896	0.5	0	0	0.00049	0	0
Deep Belief Network	0.7838	0.6898	0.5002	0.014814	0.0004	0.000748	0.00098	0.029628
GRU	0.7864	0.7362	0.5546	0.374289	0.006053	0.009064	0.011977	0.037334
CNN	0.7878	0.7278	0.5428	0.324707	0.006615	0.008376	0.010759	0.056225

Table 16 Means and standard deviations of performance metric values for all methods using the SHHS-1 dataset as a binary classes prediction problem (AHI=30 threshold) using 27 clinical features selected from the SHHS-1 cohort

	Mean				Standard deviation			
	Acc	F1	Bacc	GM	Acc	F1	Bacc	GM
Nearest Centroid	0.6926	0.7656	0.6464	0.64435	0.016895	0.012093	0.01196	0.011588
Logistic Regression	0.9272	0.8958	0.512	0.139233	0.001939	0.003487	0.009839	0.081182
SGD	0.856	0.8696	0.5878	0.496	0.015505	0.010929	0.020429	0.030263
Perceptron	0.6096	0.5978	0.5324	0.24924	0.37673	0.365107	0.025097	0.143612
Decision Tree	0.882	0.88	0.5478	0.382177	0.00469	0.003286	0.011196	0.025791
Bagging	0.8864	0.879	0.526	0.300275	0.024508	0.012602	0.014873	0.087052
Extra Trees	0.9266	0.8922	0.5012	0.025591	0.001497	0.00172	0.003429	0.051182
Hist Gradient Boosting	0.8952	0.8842	0.5306	0.304139	0.01806	0.00744	0.0145	0.077818
LDA	0.9262	0.896	0.5144	0.175143	0.003059	0.003033	0.00731	0.041504
QDA	0.9276	0.8922	0.5	0	0.00049	0.00098	0	0
BernoulliNB	0.926	0.8922	0.501	0.025591	0.001265	0.00147	0.003521	0.051182
MultinomialNB	0.9276	0.8922	0.5	0	0.00049	0.00098	0	0
GaussianNB	0.844	0.864	0.6112	0.547034	0.015875	0.010991	0.017046	0.020199
MLP	0.919	0.8968	0.5316	0.272438	0.007348	0.005706	0.015832	0.052441
SVC	0.606	0.6978	0.5448	0.531934	0.065915	0.05227	0.045195	0.052316
LinearSVC	0.9266	0.8922	0.501	0.025525	0.001855	0.00098	0.002	0.051051
NuSVC	0.5482	0.646	0.6076	0.595467	0.088375	0.078677	0.048442	0.050128
XGBoost	0.9236	0.8972	0.525	0.231705	0.002417	0.003655	0.016444	0.066035
Random Forest	0.9266	0.8932	0.5042	0.061775	0.001855	0.002786	0.006369	0.07748
AdaBoost	0.92	0.8964	0.5278	0.256053	0.004336	0.002498	0.008841	0.043494
Gradient Boosting	0.891	0.8834	0.536	0.335945	0.016876	0.010538	0.019667	0.042706
Radius Neighbors	0.9276	0.8922	0.5	0	0.00049	0.00098	0	0
Ridge	0.9276	0.8922	0.5	0	0.00049	0.00098	0	0
Passive Aggressive	0.857	0.8484	0.536	0.134066	0.140001	0.086703	0.072501	0.268132
LGBM	0.926	0.8976	0.5204	0.207359	0.003033	0.003137	0.009308	0.048422
CatBoost	0.881	0.8756	0.5528	0.342207	0.062041921	0.032879173	0.053033574	0.161027811
KNN	0.878	0.8752	0.5258	0.32535	0.006066	0.004622	0.012172	0.033427
CEM	0.9254	0.8982	0.524	0.2262	0.007736	0.009786	0.010488	0.048787
DANet	0.9232	0.8946	0.5156	0.1858	0.00581	0.0132	0.00979	0.062191
Node	0.9274	0.8922	0.5	0	0.008868	0.012576	0	0

Table 16 (continued)

	Mean				Standard deviation			
	Acc	F1	Bacc	GM	Acc	F1	Bacc	GM
GATE	0.9232	0.8958	0.5252	0.1778	0.005913	0.016339	0.028882	0.161845
TabNet	0.9252	0.893	0.5046	0.0856	0.009558	0.012474	0.004673	0.071208
Tab- Trans- former	0.9278	0.8942	0.5064	0.0878	0.008976	0.012734	0.005817	0.073295
LSTM	0.9276	0.8922	0.5	0	0.00049	0.00098	0	0
GCN	0.9268	0.8926	0.5028	0.051133	0.00147	0.001497	0.003487	0.062625
RNN	0.9276	0.8922	0.5	0	0.00049	0.00098	0	0
Deep Belief Network	0.9276	0.8922	0.5	0	0.00049	0.00098	0	0
GRU	0.9252	0.8936	0.5052	0.120351	0.004308	0.0032	0.008495	0.070037
CNN	0.9262	0.893	0.5038	0.07631	0.001166	0.000632	0.003544	0.062307

Table 17 Feature description for the SHHS-1 dataset (Zhang et al., 2018; Quan et al., 1997)

OSFP	SHHS-1	
Feature name	Feature selected?	Variable Name by NSRR
Sex	✓	gender
Age	✓	age_s1
Current smoker	✓	smokstat_s1
Former smoker		
Sedentary		
Height	✓	height
Weight	✓	weight
Cervical perimeter	✓	neck20
Abdominal perimeter	✓	waist
Systolic BP	✓	systbp
Diastolic BP	✓	diasbp
Maxillofacial profile		
BMI	✓	bmi_s1
High BP	✓	srhype
Asthma	✓	asthma15
Rhinitis		
COPD	✓	copd15
Respiratory fail		
Myocardial infarct	✓	mi15
Coronary fail		
Arrhythmias		
Stroke	✓	stroke15
Heart Fail	✓	hf15
Arteriopathy		
Gastric reflux	✓	hb02
Glaucoma		
Diabetes	✓	parrptdiab
Hypercholesterolemia	✓	chol
Hypertriglyceridemia	✓	trig
Hypo(er)thyroidism		
Depression		
Obesity		
Dysmorphology		
Restless Leg Syndrome		
Snoring		
Diurnal somnolence	✓	sleepy02
Driving drowsiness	✓	drive02
Morning fatigue	✓	funres02
Morning headache		
Memory problem		
Nocturnal perspiration	✓	sweats02
Shortness of breath on exertion	✓	sob02
Nocturia	✓	needbr02
Drowsiness accident		
Near Miss accident		
Respiratory arrest		

Table 17 (continued)

OSFP	SHHS-1	
Feature name	Feature selected?	Variable Name by NSRR
Epworth scale	✓	ess_s1
Pichots scale		
Depression scale		

A feature set of 27 variables for the SHHS-1 is selected based on the selected feature set from the OSFP data, and features with a missing rate of more than 10% are also removed. After excluding patients with missing values in selected features, 4,218 patients are included in our experiment

consistent. First, OSA severity assessments are hard with low prediction accuracy for all tasks regardless used models. Second, recent DL methods like DANet, CEM or NODE and ensemble methods like CatBoost or LGBM consistently performs better than the others.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10994-025-06937-4>.

Acknowledgements We would like to give special thanks to anonymous reviewers for their valuable comments that have helped to significantly enhance the quality of the paper. The Sleep Heart Health Study (SHHS) was supported by National Heart, Lung, and Blood Institute cooperative agreements U01HL53916 (University of California, Davis), U01HL53931 (New York University), U01HL53934 (University of Minnesota), U01HL53937 and U01HL64360 (Johns Hopkins University), U01HL53938 (University of Arizona), U01HL53940 (University of Washington), U01HL53941 (Boston University), and U01HL63463 (Case Western Reserve University). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002). NSRR R24 HL114473: NHLBI National Sleep Research Resource.

Author Contributions HTKC, ADB, AJL, DTP, PND, HVT, TLH, XHN, STM: Conceptualization, Data curation, Methodology, Software, Writing – original draft, Writing – review & editing. JM, HTKC, ZZ, XHN, TLH, YA, PL, HPN, STM: Conceptualization, Writing – original draft, Writing – review & editing. AJL, XS, HV, SB, JLP, STM: Conceptualization, Data curation, Methodology, Supervision, Writing – review & editing. All authors reviewed the manuscript.

Funding JLP and SB are supported by the French National Research Agency (ANR-15-IDEX-02 and ANR-24-CE36-3468-01), MIAI artificial intelligence chair (ANR-23-IACL-0006) and chair E-Santé from Grenoble Alpes University Foundation. Zichi Zhang is supported by the China Scholarship Council (CSC) under Grant No. 202410080008. Yimeng An is supported by Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX23 0404).

Data Availability Observatoire Sommeil de la Federation de Pneumologie” (OSFP).

Declarations

Conflict of interest The authors declare no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are

included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Agrawal, P., Kumari, R., & Das, P., et al. (2025). A novel deep-learning method for obstructive sleep apnea detection from single channel photoplethysmography. In: *ISCAS*, pp. 1–5.
- Ahbab, S., Ataoglu, H. E., Tuna, M., et al. (2013). Neck circumference, metabolic syndrome and obstructive sleep apnea syndrome; evaluation of possible linkage. *Medical Science Monitor*, 19, 111.
- Alarcón, Á. S., Gaiduk, M., Madrid, N. M., et al. (2024). Classification of the sleep-wake state through the development of a deep learning model. *Procedia Computer Science*, 246, 4636–4645.
- Álvarez, D., Arroyo, C. A., de Frutos, J. F., et al. (2020). Assessment of nocturnal autonomic cardiac imbalance in positional obstructive sleep apnea: A multiscale nonlinear approach. *Entropy*, 22(12), 1404.
- Alvarez, D., Hornero, R., Marcos, J. V., et al. (2010). Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis. *IEEE Transactions on Biomedical Engineering*, 57(12), 2816–2824.
- Alzubaidi, L., Zhang, J., Humaidi, A. J., et al. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of big Data*, 8(1), 53.
- An, H. J., Baek, S. H., Kim, S. W., et al. (2020). Clustering-based characterization of clinical phenotypes in obstructive sleep Apnoea using severity, obesity, and craniofacial pattern. *European Journal of Orthodontics*, 42(1), 93–100.
- Appleton, S., Gill, T., Taylor, A., et al. (2018). Influence of gender on associations of obstructive sleep apnea symptoms with chronic conditions and quality of life. *International Journal of Environmental Research and Public Health*, 15(5), 930.
- Arik, SÖ., & Pfister, T. (2021). Tabnet: Attentive interpretable tabular learning. In: *AAAI*, pp. 6679–6687.
- Arnardottir, E. S., Bjornsdottir, E., Olafsdottir, K. A., et al. (2015). Obstructive sleep apnoea in the general population: Highly prevalent but minimal symptoms. *European Respiratory Journal*, 47(1), 194–202.
- Azimi, H., Xi, P., Bouchard, M., et al. (2020). Machine learning-based automatic detection of central sleep apnea events from a pressure sensitive mat. *IEEE Access*, 8, 173428–173439.
- Bailly, S., Destors, M., Grillet, Y., et al. (2016). Obstructive sleep apnea: A cluster analysis at time of diagnosis. *PLoS ONE*, 11(6), Article e0157318.
- Bailly, S., Grote, L., Hedner, J., et al. (2021). Clusters of sleep apnoea phenotypes: A large pan-European study from the European sleep apnoea database (Esada). *Respirology*, 26(4), 378–387.
- Bark, B., Nam, B., & Kim, I. Y. (2023). Selanet: Decision-assisting selective sleep apnea detection based on confidence score. *BMC Medical Informatics and Decision Making*, 23(1), 190.
- Behar, J., Roebuck, A., Shahid, M., et al. (2014). Sleepap: An automated obstructive sleep apnoea screening application for smartphones. *IEEE Journal of Biomedical and Health Informatics*, 19(1), 325–331.
- Ben Sason, Y., Oksenberg, A., Sobel, J. A., et al. (2023). Characteristics of patients with positional osa according to ethnicity and the identification of a novel phenotype-lateral positional patients: A multi-ethnic study of atherosclerosis (mesa) study. *Journal of Clinical Sleep Medicine*, 19(3), 529–538.
- Benjafield, A. V., Ayas, N. T., Eastwood, P. R., et al. (2019). Estimation of the global prevalence and burden of obstructive sleep apnoea: A literature-based analysis. *The Lancet Respiratory Medicine*, 7(8), 687–698. [https://doi.org/10.1016/S2213-2600\(19\)30198-5](https://doi.org/10.1016/S2213-2600(19)30198-5). epub 2019 Jul 9.
- Bernardini, A., Brunello, A., Gigli, G. L., et al. (2021). Aiosa: An approach to the automatic identification of obstructive sleep apnea events based on deep learning. *Artificial Intelligence in Medicine*, 118, Article 102133.
- Bonsignore, M. R., Baiamonte, P., Mazzuca, E., et al. (2019). Obstructive sleep apnea and comorbidities: A dangerous liaison. *Multidisciplinary Respiratory Medicine*, 14(1), 8.
- Borisov, V., Leemann, T., Seßler, K., et al. (2022). Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6), 7499–7519.
- Botelho, M.C., Trancoso, I., & Abad, A., et al. (2019). Speech as a biomarker for obstructive sleep apnea detection. In: *ICASSP*, pp. 5851–5855.
- Burgos, A., Goni, A., Illarramendi, A., et al. (2009). Real-time detection of apneas on a pda. *IEEE Transactions on Information Technology in Biomedicine*, 14(4), 995–1002.
- Camci, B., Kahveci, A.Y., & Amrich, B., et al. (2017). Sleep apnea detection via smart phones. In: *SIU*, pp. 1–4.

- Chawla, N. V., Bowyer, K. W., Hall, L. O., et al. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
- Chen, J., Liao, K., Wan, Y., et al. (2022). Danets: Deep abstract networks for tabular data classification and regression. In: AAAI, pp. 3930–3938.
- Chen, L., Zhang, X., & Song, C. (2014). An automatic screening approach for obstructive sleep apnea diagnosis based on single-lead electrocardiogram. *IEEE Transactions on Automation Science and Engineering*, 12(1), 106–115.
- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In: KDD, pp. 785–794.
- Chen, X., Wang, R., Zee, P., et al. (2015). Racial/ethnic differences in sleep disturbances: The multi-ethnic study of atherosclerosis (mesa). *Sleep*, 38(6), 877–888.
- Chen, Y. H., Keller, J. K., Kang, J. H., et al. (2013). Obstructive sleep apnea and the subsequent risk of depressive disorder: A population-based follow-up study. *Journal of Clinical Sleep Medicine*, 9(5), 417–423.
- Cheng, W. J., Finnsson, E., Arnardóttir, E., et al. (2023). Relationship between symptom profiles and endotypes among patients with obstructive sleep apnea: A latent class analysis. *Annals of the American Thoracic Society*, 20(9), 1337–1344.
- Chiang, C. L., Chen, Y. T., Wang, K. L., et al. (2017). Comorbidities and risk of mortality in patients with sleep apnea. *Annals of Medicine*, 49(5), 377–383.
- Chung, F., Abdullah, H. R., & Liao, P. (2016). Stop-bang questionnaire: A practical approach to screen for obstructive sleep apnea. *Chest*, 149(3), 631–638.
- Cochran, W. G. (1952). The χ^2 test of goodness of fit. *Annals of Mathematical Statistics*, pp. 315–345.
- De Chazal, P., Sutherland, K., & Cistulli, P. A. (2020). Advanced polysomnographic analysis for osa: A pathway to personalized management? *Respirology*, 25(3), 251–258.
- Deviaene, M., Testelmans, D., Buysse, B., et al. (2018). Automatic screening of sleep apnea patients based on the spo 2 signal. *IEEE Journal of Biomedical and Health Informatics*, 23(2), 607–617.
- Engleman, H. M., & Douglas, N. J. (2004). Sleep.4: Sleepiness, cognitive function, and quality of life in obstructive sleep apnoea/hypopnoea syndrome. *Thorax*, 59(7), 618–622.
- Farhangfar, A., Kurgan, L., & Dy, J. (2008). Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition*, 41(12), 3692–3705.
- Ferreira-Santos, D., & Rodrigues, P. P. (2018). Finding groups in obstructive sleep apnea patients: A categorical cluster analysis. In: CBMS, pp. 387–392.
- Ferreira-Santos, D., & Rodrigues, P. P. (2021). Enhancing obstructive sleep apnea diagnosis with screening through disease phenotypes: Algorithm development and validation. *JMIR Medical Informatics*, 9(6), Article e25124.
- Ferreira-Santos, D., & Rodrigues, P. P. (2023). Obstructive sleep apnea: A categorical cluster analysis and visualization. *Pulmonology*, 29(3), 207–213.
- Fietze, I., Laharnar, N., Obst, A., et al. (2019). Prevalence and association analysis of obstructive sleep apnea with gender and age differences—results of ship-trend. *Journal of Sleep Research*, 28(5), Article e12770.
- Franklin, K. A., & Lindberg, E. (2015). Obstructive sleep apnea is a common disorder in the population—A review on the epidemiology of sleep apnea. *Journal of Thoracic Disease*, 7(8), 1311.
- Ghassemi, M. M., Moody, B. E., Lehman, L. W. H., et al. (2018). You snooze, you win: The physionet/computing in cardiology challenge 2018. In: CinC, pp. 1–4.
- Ghorvei, M., Karhu, T., Hietakoste, S., et al. (2025). A comparative analysis of unsupervised machine-learning methods in psg-related phenotyping. *Journal of Sleep Research*, 34(3), Article e14349.
- Goldberger, A. L., Amaral, L. A., Glass, L., et al. (2000). Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23), e215–e220.
- Gottlieb, D. J., & Punjabi, N. M. (2020). Diagnosis and management of obstructive sleep apnea: A review. *JAMA*, 323(14), 1389–1400.
- Gou, J., Yu, B., Maybank, S. J., et al. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6), 1789–1819.
- Gutiérrez-Tobal, G. C., Álvarez, D., Crespo, A., et al. (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.
- Hafezi, M., Montazeri, N., Saha, S., et al. (2020). Sleep apnea severity estimation from tracheal movements using a deep learning model. *IEEE Access*, 8, 22641–22649.
- Haidar, R., Koprinska, I., & Jeffries, B. (2017). Sleep apnea event detection from nasal airflow using convolutional neural networks. In: ICONIP, pp. 819–827.
- Han, H., & Oh, J. (2023). Application of various machine learning techniques to predict obstructive sleep apnea syndrome severity. *Scientific Reports*, 13(1), 6379.
- Hassan, A. R. (2016). Computer-aided obstructive sleep apnea detection using normal inverse Gaussian parameters and adaptive boosting. *Biomedical Signal Processing and Control*, 29, 22–30.

- Hassan, A. R., & Haque, M. A. (2017). An expert system for automated identification of obstructive sleep apnea from single-lead ecg using random under sampling boosting. *Neurocomputing*, 235, 122–130.
- He, H., Bai, Y., Garcia, E. A., et al. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: *Proc. 2008 IEEE Int. Jt. Conf. Neural Netw.* (IEEE World Congr. Comput. Intell.), pp. 1322–1328.
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284.
- Hemrajani, P., Dhaka, V. S., Rani, G., et al. (2023). Efficient deep learning based hybrid model to detect obstructive sleep apnea. *Sensors*, 23(10), 4692.
- Hemrajani, P., Dhaka, V. S., Rani, G., et al. (2025). Integrating physiological signals for enhanced sleep apnea diagnosis with sleepnet. *Scientific Reports*, 15(1), 31715.
- Heneghan, C. (2011). *St. Vincent's university hospital/university college Dublin sleep apnea database*. Vincent's university hospital/University College Dublin sleep apnea database.
- Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7), 1527–1554.
- Hou, H., Zhao, Y., Yu, W., et al. (2018). Association of obstructive sleep apnea with hypertension: A systematic review and meta-analysis. *Journal of Global Health*, 8(1), Article 010405.
- Howarth, T., Hedger, J., Chen, W., et al. (2025). General practice sleep scale-the“gpss”—a proposed new tool for use in general practice for risk assessment of obstructive sleep apnoea. *Sleep Medicine*, 125, 168–176.
- Huang, W. C., Lee, P. L., Liu, Y. T., et al. (2020). Support vector machine prediction of obstructive sleep apnea in a large-scale Chinese clinical sample. *Sleep*, 43(7), zsz295.
- Huang, X., Khetan, A., Cvitkovic, M., et al. (2020b). *Tabtransformer: Tabular data modeling using contextual embeddings*. arXiv preprint [arXiv:2012.06678](https://arxiv.org/abs/2012.06678)
- Hu, S., Wang, Y., Liu, J., et al. (2025). Xsleepfusion: A dual-stage information bottleneck fusion framework for interpretable multimodal sleep analysis. *Information Fusion*, 123, Article 103275.
- Islam, S. M., Mahmood, H., Al-Jumaily, A. A., et al. (2018). Deep learning of facial depth maps for obstructive sleep apnea prediction. In: *iCMLDE*, pp. 154–157.
- Ivanko, K., Ivanushkina, N., Rykhalska, A. (2020). Identifying episodes of sleep apnea in ecg by machine learning methods. In: *ELNANO*, pp. 588–593.
- Javaheri, S., Barbe, F., Campos-Rodriguez, F., et al. (2017). Sleep apnea: Types, mechanisms, and clinical cardiovascular consequences. *Journal of the American College of Cardiology*, 69(7), 841–858.
- Javaid, A. Q., Noble, C. M., Rosenberg, R., et al. (2015). Towards sleep apnea screening with an under-the-mattress ir-uwv radar using machine learning. In: *ICMLA*, pp. 837–842.
- Jayawardhana, M., de Chazal, P. (2017). Enhanced detection of sleep apnoea using heart-rate, respiration effort and oxygen saturation derived from a photoplethysmography sensor. In: *EMBC*, pp. 121–124.
- Jeon, Y., Heo, K., & Kang, S. J. (2020). Real-time sleep apnea diagnosis method using wearable device without external sensors. In: *Proc. 2020 IEEE Int. Conf. Pervasive Comput. Commun. Workshops*, pp. 1–5.
- Jiang, D., Ma, Y., & Wang, Y. (2018). A multi-scale parallel convolutional neural network for automatic sleep apnea detection using single-channel ecg signals. In: *CISP-bMEI*, pp. 1–5.
- Johns, M. W. (1991). A new method for measuring daytime sleepiness: The epworth sleepiness scale. *Sleep*, 14(6), 540–545.
- Kapur, V. K., Auckley, D. H., Chowdhuri, S., et al. (2017). Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: An American academy of sleep medicine clinical practice guideline. *Journal of Clinical Sleep Medicine*, 13(3), 479–504.
- Ke, G., Meng, Q., Finley, T., et al. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *NeurIPS* 30.
- Keenan, B. T., Kim, J., Singh, B., et al. (2018). Recognizable clinical subtypes of obstructive sleep apnea across international sleep centers: A cluster analysis. *Sleep*, 41(3), zsx214.
- Kelleher, J. D., Mac Namee, B., & D'arcy, A. (2020). *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies*. MIT press.
- Keshavarz, Z., Rezaee, R., Nasiri, M., et al. (2020) Obstructive sleep apnea: A prediction model using supervised machine learning method. In: *Stud. Health Technol. Inform.* pp. 387–390.
- Khan, A., Basit, M. S., Farooq, O., et al. (2023). Mitigating the class imbalance effect in sleep apnea classification. In: *REEDCON*, pp. 530–535.
- Khandoker, A. H., Gubbi, J., & Palaniswami, M. (2009). Automated scoring of obstructive sleep apnea and hypopnea events using short-term electrocardiogram recordings. *IEEE Transactions on Information Technology in Biomedicine*, 13(6), 1057–1067.
- Khandoker, A. H., Palaniswami, M., & Karmakar, C. K. (2008). Support vector machines for automated recognition of obstructive sleep apnea syndrome from ecg recordings. *IEEE Transactions on Information Technology in Biomedicine*, 13(1), 37–48.

- Kim, H., Yun, C. H., Thomas, R. J., et al. (2013). Obstructive sleep apnea as a risk factor for cerebral white matter change in a middle-aged and older general population. *Sleep*, 36(5), 709–715.
- Kim, J. W., Won, T. B., Rhee, C. S., et al. (2020). Polysomnographic phenotyping of obstructive sleep apnea and its implications in mortality in Korea. *Scientific Reports*, 10(1), 13207.
- Kim, T., Kim, J. W., & Lee, K. (2018). Detection of sleep disordered breathing severity using acoustic biomarker and machine learning techniques. *Biomedical Engineering Online*, 17(1), 16.
- Kim, Y. J., Jeon, J. S., Cho, S. E., et al. (2021). Prediction models for obstructive sleep apnea in Korean adults using machine learning techniques. *Diagnostics (Basel, Switzerland)*, 11(4), 612.
- Knauert, M., Naik, S., Gillespie, M. B., et al. (2015). Clinical consequences and economic costs of untreated obstructive sleep apnea syndrome. *World Journal Otolaryngology-Head & Neck Surgery*, 1(01), 17–27.
- Kuo, N. Y., Tsai, H. J., Tsai, S. J., et al. (2024). Efficient screening in obstructive sleep apnea using sequential machine learning models, questionnaires, and pulse oximetry signals: Mixed methods study. *Journal of Medical Internet Research*, 26, Article e51615.
- Kwon, H. B., Son, D., Lee, D., et al. (2021). Hybrid cnn-lstm network for real-time apnea-hypopnea event detection based on ir-uwb radar. *IEEE Access*, 10, 17556–17564.
- Lacedonia, D., Carpagnano, G. E., Sabato, R., et al. (2016). Characterization of obstructive sleep apnea-hypopnea syndrome (osa) population by means of cluster analysis. *Journal of Sleep Research*, 25(6), 724–730.
- Lakhan, P., Dithapron, A., Banluesombatkul, N., et al. (2018). Deep neural networks with weighted averaged overnight airflow features for sleep apnea-hypopnea severity classification. In: TENCON, pp. 0441–0445.
- Lee, Y. H., Jeon, S., Auh, Q. S., et al. (2024). Automatic prediction of obstructive sleep apnea in patients with temporomandibular disorder based on multidata and machine learning. *Scientific Reports*, 14(1), 19362.
- Liang, Z. (2023). Novel method combining multiscale attention entropy of overnight blood oxygen level and machine learning for easy sleep apnea screening. *Digital Health*, 9, 20552076231211550.
- Li, K., Pan, W., Li, Y., et al. (2018). A method to detect sleep apnea based on deep neural network and hidden Markov model using single-lead ecg signal. *Neurocomputing*, 294, 94–101.
- Li, Y., Miao, Y., Tan, J., et al. (2023). Association of modifiable risk factors with obstructive sleep apnea: A mendelian randomization study. *Aging (Albany NY)*, 15(23), 14039.
- Liu, J., Li, Q., Xin, Y., et al. (2020). Obstructive sleep apnea detection using sleep architecture. In: ICMA, pp. 255–260.
- Mahmud, T., Khan, I. A., Mahmud, T. I., et al. (2021). Sleep apnea detection from variational mode decomposed ecg signal using a hybrid cnn-bilstm. *IEEE Access*, 9, 102355–102367.
- Mai, S. T., Amer-Yahia, S., Bailly, S., et al. (2018). Evolutionary active constrained clustering for obstructive sleep apnea analysis. *Data Science and Engineering*, 3(4), 359–378.
- Ma, E. Y., Kim, J. W., Lee, Y., et al. (2021). Combined unsupervised-supervised machine learning for phenotyping complex diseases with its application to obstructive sleep apnea. *Scientific Reports*, 11(1), 4457.
- Mayo Clinic. (2020). Sleep apnea? <https://www.mayoclinic.org/diseases-conditions/sleep-apnea/symptom-s-causes/syc-20377631>, cited 2022 Aug 8
- Mencar, C., Gallo, C., Mantero, M., et al. (2020). Application of machine learning to predict obstructive sleep apnea syndrome severity. *Health Informatics Journal*, 26(1), 298–317.
- Montazeri Ghahjaverestan, N., Aguiar, C., Hummel, R., et al. (2023). Sleep apnea detection by tracheal motion and sound, and oximetry via application of deep neural networks. *Nature and Science of Sleep*. <https://doi.org/10.2147/NSS.S397196>
- Moradhasel, B., Sheikhan, A., Aloosh, O., et al. (2023). Spectrogram classification of patient chin electromyography based on deep learning: A novel method for accurate diagnosis obstructive sleep apnea. *Biomedical Signal Processing and Control*, 79, Article 104215.
- Morillo, D. S., & Gross, N. (2013). Probabilistic neural network approach for the detection of sahs from overnight pulse oximetry. *Medical & Biological Engineering & Computing*, 51(3), 305–315.
- Mostafa, S. S., Mendonça, F., Ravelo-García, G. A., et al. (2019). A systematic review of detecting sleep apnea using deep learning. *Sensors*, 19(22), 4934.
- National Heart Lung, & Blood Institute (2022). What is sleep apnea? <https://www.nhlbi.nih.gov/health/sleep-apnea>, cited 2022 Aug 8
- Nikkonen, S., Afara, I. O., Leppänen, T., et al. (2019). Artificial neural network analysis of the oxygen saturation signal enables accurate diagnostics of sleep apnea. *Scientific Reports*, 9(1), 13200.
- Nishad, A., Pachori, R. B., & Acharya, U. R. (2024). Application of tqwt based filter-bank for sleep apnea screening using ecg signals. *Journal of Ambient Intelligence and Humanized Computing*, 15(1), 893–904.
- Olsen, M., Mignot, E., Jennum, P. J., et al. (2020). Robust, ecg-based detection of sleep-disordered breathing in large population-based cohorts. *Sleep*, 43(5), zsz276.

- Padovano, D., Martinez-Rodrigo, A., Pastor, J. M., et al. (2025). Deep learning and recurrence information analysis for the automatic detection of obstructive sleep apnea. *Applied Sciences*, 15(1), 433.
- Park, K., Hong, J., Lee, W., et al. (2025). Distillsleep: Real-time, on-device, interpretable sleep staging from single-channel eeg. *Sleep*, p. zsaf240.
- Penzel, T., Kantelhardt, J. W., Bartsch, R. P., et al. (2016). Modulations of heart rate, ECG, and cardio-respiratory coupling observed in polysomnography. *Frontiers in Physiology*, 7, 460.
- Penzel, T., Moody, G.B., Mark, R.G., et al. (2000). The apnea-ecg database. In: *Computers in cardiology 2000*. Vol. 27 (Cat. 00CH37163), pp. 255–258.
- Peppard, P. E., Young, T., Palta, M., et al. (2000). Prospective study of the association between sleep-disordered breathing and hypertension. *The New England Journal of Medicine*, 342(19), 1378–1384.
- Petrenko, A. (2020). Breathmonitor: Ai sleep apnea mobile detector. In: SAIC, pp. 201–208.
- Pichot, P., & Brun, J. P. (1984). Brief self-evaluation questionnaire for depressive, asthenic and anxious dimensions. *Annales Médico-Psychologiques*, 142, 862–865.
- Pinto, J. A., Ribeiro, D. K., da Silva Cavallini, A. F., et al. (2016). Comorbidities associated with obstructive sleep apnea: A retrospective study. *International Archives of Otorhinolaryngology*, 20(02), 145–150.
- Prokhorenkova, L., Gusev, G., Vorobev, A., et al. (2018). Catboost: unbiased boosting with categorical features. *NeurIPS* 31.
- Punjabi, N. M. (2008). The epidemiology of adult obstructive sleep apnea. *Proceedings of the American Thoracic Society*, 5(2), 136–143.
- Punjabi, N. M., Caffo, B. S., Goodwin, J. L., et al. (2009). Sleep-disordered breathing and mortality: A prospective cohort study. *PLoS Medicine*, 6(8), Article e1000132.
- Qiu, X., Wei, Y., Tan, X., et al. (2026). Mimar-osa: Enhancing obstructive sleep apnea diagnosis through multimodal data integration and missing modality reconstruction. *Pattern Recognition*, 169, Article 111917.
- Quan, S. F., Howard, B. V., Iber, C., et al. (1997). The sleep heart health study: Design, rationale, and methods. *Sleep*, 20(12), 1077–1085.
- Quintana-Gallego, E., Carmona-Bernal, C., Capote, F., et al. (2004). Gender differences in obstructive sleep apnea syndrome: A clinical study of 1166 patients. *Respiratory Medicine*, 98(10), 984–989.
- Ramachandran, A., & Karuppiiah, A. (2021). A survey on recent advances in machine learning based sleep apnea detection systems. *Healthcare (Basel)*. <https://doi.org/10.3390/healthcare9070914>
- Ramesh, J., Keeran, N., Sagahyroon, A., et al. (2021). Towards validating the effectiveness of obstructive sleep apnea classification from electronic health records using machine learning. *Healthcare (Basel)*. <https://doi.org/10.3390/healthcare9111450>
- Randerath, W., Bassetti, C. L., Bonsignore, M. R., et al. (2018). Challenges and perspectives in obstructive sleep apnoea: Report by an ad hoc working group of the sleep disordered breathing group of the European Respiratory Society and the European Sleep Research Society. *European Respiratory Journal*, 52(3), Article 1702616.
- Razjouyan, J., Hanania, N. A., Nowakowski, S., et al. (2024). Identification of sleep phenotypes in copd using machine learning-based cluster analysis. *Respiratory Medicine*, 227, Article 107641.
- Rodrigues Jr J. F., Pepin, J. L., Goeuriot, L., et al. (2020). An extensive investigation of machine learning techniques for sleep apnea screening. In: CIKM, pp. 2709–2716.
- Romero-Corral, A., Caples, S. M., Lopez-Jimenez, F., et al. (2010). Interactions between obesity and obstructive sleep apnea: Implications for treatment. *Chest*, 137(3), 711–719.
- Rundo, J. V. (2019). Obstructive sleep apnea basics. *Cleveland Clinic Journal of Medicine*, 86(9 Suppl 1), 2–9.
- Sateia, M. J. (2014). International classification of sleep disorders. *Chest*, 146(5), 1387–1394.
- Schütz, S. G., Lisabeth, L. D., Shafie-Khorassani, F., et al. (2019). Clinical phenotypes of obstructive sleep apnea after ischemic stroke: A cluster analysis. *Sleep Medicine*, 60, 178–181.
- Semelka, M., Wilson, J., & Floyd, R. (2016). Diagnosis and treatment of obstructive sleep apnea in adults. *American Family Physician*, 94(5), 355–360.
- Shahid, MLUR., Chitiboi, T., Ivanovska, T., et al. (2015). Automatic pharynx segmentation from mri data for obstructive sleep apnea analysis. In: VISAPP, pp. 599–608.
- Shanmugham, A., Srivatsa, B. V. A., Gopikrishnan, K., et al. (2021). Sleep apnea detection using resnet. In: ICCNT, pp. 1–4.
- Sharma, H., & Sharma, K. (2016). An algorithm for sleep apnea detection from single-lead eeg using hermite basis functions. *Computers in Biology and Medicine*, 77, 116–124.
- Sharma, M., Kumbhani, D., Tiwari, J., et al. (2022). Automated detection of obstructive sleep apnea in more than 8000 subjects using frequency optimized orthogonal wavelet filter bank with respiratory and oximetry signals. *Computers in Biology and Medicine*, 144, Article 105364.
- Sharma, P., Jalali, A., Majmudar, M., et al. (2022b). Deep-learning based sleep apnea detection using spo2 and pulse rate. In: EMBC, pp. 2611–2614.

- Sobek, A. (1999). Final results of incorporating an operational formal method into a software engineering curriculum. In: FIE, pp. 13A3–18.
- Song, C., Liu, K., Zhang, X., et al. (2015). An obstructive sleep apnea detection approach using a discriminative hidden Markov model from ecg signals. *IEEE Transactions on Biomedical Engineering*, 63(7), 1532–1542.
- Stanford Health Care (SHC) - Stanford Medical Center (2017) Diagnosis. <https://stanfordhealthcare.org/medical-conditions/sleep/obstructive-sleep-apnea/diagnosis.html>, cited 2022 Aug 8
- Sunil Kumar, T., & Kanhangad, V. (2017). Automated obstructive sleep apnoea detection using symmetrically weighted local binary patterns. *Electronics Letters*, 53(4), 212–214.
- Sutherland, K., Almeida, F. R., Kim, T., et al. (2021). Treatment usage patterns of oral appliances for obstructive sleep apnea over the first 60 days: A cluster analysis. *Journal of Clinical Sleep Medicine*, 17(9), 1785–1792.
- Tan, H.L., Gozal, D., Kheirandish-Gozal, L. (2013). Obstructive sleep apnea in children: A critical update. *Nature and Science of Sleep*, pp. 109–123.
- Teran-Santos, J., Jimenez-Gomez, A., & Cordero-Guevara, J. (1999). The association between sleep apnea and the risk of traffic accidents. *The New England Journal of Medicine*, 340(11), 847–851.
- Tsuchiya, M., Lowe, A. A., Pae, E. K., et al. (1992). Obstructive sleep apnea subtypes by cluster analysis. *American Journal of Orthodontics and Dentofacial Orthopedics*, 101(6), 533–542.
- Urtanasan, E., Park, J. U., Joo, E. Y., et al. (2018). Automated detection of obstructive sleep apnea events from a single-lead electrocardiogram using a convolutional neural network. *Journal of Medical Systems*, 42(6), 104.
- Ustun, B., Westover, M. B., Rudin, C., et al. (2016). Clinical prediction models for sleep apnea: The importance of medical history over symptoms. *Journal of Clinical Sleep Medicine*, 12(2), 161–168.
- Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45, 1–67.
- Van Der Aar, J. F., Van Den Ende, D. A., Fonseca, P., et al. (2024). Deep transfer learning for automated single-lead EEG sleep staging with channel and population mismatches. *Frontiers in Physiology*, 14, 1287342.
- Van Overstraeten, C., Andreozzi, F., Youssef, S. B., et al. (2021). Obstructive sleep apnea syndrome phenotyping by cluster analysis: Typical sleepy, obese middle-aged men with desaturating events are a minority of patients in a multi-ethnic cohort of 33% women. *Current Medical Science*, 41(4), 729–736.
- Van Steenkiste, T., Groenendaal, W., Deschrijver, D., et al. (2018). Automated sleep apnea detection in raw respiratory signals using long short-term memory neural networks. *IEEE Journal of Biomedical and Health Informatics*, 23(6), 2354–2364.
- Varon, C., Caicedo, A., Testelmans, D., et al. (2015). A novel algorithm for the automatic detection of sleep apnea from single-lead ECG. *IEEE Transactions on Biomedical Engineering*, 62(9), 2269–2278.
- Varon, C., Testelmans, D., Buyse, B., et al. (2013). Sleep apnea classification using least-squares support vector machines on single lead ecg. In: EMBC, pp. 5029–5032.
- Vavougios, G. D., Natsios, G., Pastaka, C., et al. (2016). Phenotypes of comorbidity in OSAS patients: Combining categorical principal component analysis with cluster analysis. *Journal of Sleep Research*, 25(1), 31–38.
- Venkatnarayan, K., Krishnaswamy, U. M., Rajamuri, N. K. R., et al. (2023). Identifying phenotypes of obstructive sleep apnea using cluster analysis. *Sleep and Breathing*, 27(3), 879–886.
- Viswabhargav, C. S., Tripathy, R., & Acharya, U. R. (2019). Automated detection of sleep apnea using sparse residual entropy features with various dictionaries extracted from heart rate and edr signals. *Computers in Biology and Medicine*, 108, 20–30.
- Wang, D., Zhou, Y., Chen, R., et al. (2023). The relationship between obstructive sleep apnea and asthma severity and vice versa: A systematic review and meta-analysis. *European Journal of Medical Research*, 28(1), 139.
- Wang, H., Liang, Q., Hancock, J. T., et al. (2024). Feature selection strategies: A comparative analysis of shap-value and importance-based methods. *Journal of Big Data*, 11(1), 44.
- Wang, J., Wang, Y., Feng, J., et al. (2013). Complex sleep apnea syndrome. *Patient prefer adherence*. pp. 633–641.
- Wu, M. F., Huang, W. C., Juang, C. F., et al. (2016). A new method for self-estimation of the severity of obstructive sleep apnea using easily available measurements and neural fuzzy evaluation system. *IEEE Journal of Biomedical and Health Informatics*, 21(6), 1524–1532.
- Xie, B., & Minn, H. (2012). Real-time sleep apnea detection by classifier combination. *IEEE Transactions on Information Technology in Biomedicine*, 16(3), 469–477.
- Xu, S., Faust, O., Seoni, S., et al. (2022). A review of automated sleep disorder detection. *Computers in Biology and Medicine*, 150, Article 106100.

- Ye, L., Pien, G. W., Ratcliffe, S. J., et al. (2014). The different clinical faces of obstructive sleep apnoea: A cluster analysis. *European Respiratory Journal*, 44(6), 1600–1607.
- Yılmaz, B., Asyalı, M. H., Arıkan, E., et al. (2010). Sleep stage and obstructive apnea epoch classification using single-lead ecg. *Biomedical Engineering Online*, 9(1), 39.
- Young, T., Peppard, P. E., & Taheri, S. (2005). Excess weight and sleep-disordered breathing. *Journal of Applied Physiology* (1985), 99(4), 1592–1599.
- Young, T., Skatrud, J., & Peppard, P. E. (2004). Risk factors for obstructive sleep apnea in adults. *JAMA*, 291(16), 2013–2016.
- Yousefi, Z., Bakar, F.C., de Zambotti, M., et al. (2025). Deeparousal-net: A multi-block recurrent deep learning model for proactive forecasting of non-apneic arousals from multichannel psg. *IEEE Transactions on Biomedical Engineering*.
- Yue, H., Lin, Y., Wu, Y., et al. (2021). Deep learning for diagnosis and classification of obstructive sleep apnea: A nasal airflow-based multi-resolution residual network. *Nature and Science of Sleep*. <https://doi.org/10.2147/NSS.S297856>
- Zarei, A., & Asl, B. M. (2020). Performance evaluation of the spectral autocorrelation function and autoregressive models for automated sleep apnea detection using single-lead ecg signal. *Computer Methods and Programs in Biomedicine*, 195, Article 105626.
- Zhang, D. H., Zhou, J., Wickens, J. D., et al. (2025a). Apnea and hypopnea event detection using eeg, emg, and sleep stage labels in a cohort of patients with suspected sleep apnea. *Biomedical Signal Processing and Control*, 105, Article 107628.
- Zhang, G. Q., Cui, L., Mueller, R., et al. (2018). The national sleep research resource: Towards a sleep data commons. *Journal of the American Medical Informatics Association : JAMIA*, 25(10), 1351–1358.
- Zhang, X. L., Zhang, L., Li, Y. M., et al. (2022). Multidimensional assessment and cluster analysis for osa phenotyping. *Journal of Clinical Sleep Medicine : JCSM*, 18(7), 1779–1788.
- Zhang, Y., Zhou, L., Zhu, S., et al. (2025b). Deep learning for obstructive sleep apnea detection and severity assessment: A multimodal signals fusion multiscale transformer model. *Nature and Science of Sleep*. <https://doi.org/10.2147/NSS.S492806>
- Zhuang, F., Qi, Z., Duan, K., et al. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1), 43–76.
- Zhuang, Z., Wang, F., Yang, X., et al. (2022). Accurate contactless sleep apnea detection framework with signal processing and machine learning methods. *Methods*, 205, 167–178.
- Zinchuk, A., & Yaggi, H. K. (2019). Sleep apnea heterogeneity, phenotypes, and cardiovascular risk. *Implications for trial design and precision sleep medicine*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Huynh Thi Khanh Chi¹ · Amonae Dabbs-Brown² · Anna Jurek-Loughrey¹ · James Mulhail¹ · Tuan Dung Pham¹ · Ngoc Phu Doan¹ · Viet Hung Tran¹ · Zichi Zhang¹ · Xuan Hoang Nguyen⁵ · Yimeng An^{1,3} · Peixin Li^{1,3} · Phi Hung Nguyen¹ · Thi Linh Hoang⁶ · Xinming Shi¹ · Hans Vandierendonck¹ · Sebastien Bailly⁴ · Jean-Louis Pépin⁴ · Thai Son Mai¹

✉ Anna Jurek-Loughrey
a.jurek@qub.ac.uk

✉ Thai Son Mai
thaison.mai@qub.ac.uk

Huynh Thi Khanh Chi
thuynh01@qub.ac.uk

Sebastien Bailly
sbailly@chu-grenoble.fr

Jean-Louis Pépin
jpepin@chu-grenoble.fr

- ¹ Queen's University Belfast, Belfast, UK
- ² Emory University, Atlanta, USA
- ³ Nanjing University of Aeronautics and Astronautics, Nanjing, China
- ⁴ University of Grenoble Alpes, Grenoble, France
- ⁵ Institut Polytechnique de Paris, Palaiseau, France
- ⁶ Singapore Management University, Singapore, Singapore