1) What is the main claim of the paper? Why is this an important contribution to the machine learning literature?

Obstructive Sleep Apnoea (OSA) is a most common sleep disorder characterized by the collapsing of the upper airways during sleep, causing restrictions in airflow and periodic episodes of apnoea and hypopnea. It consequently leads to many significant adverse effects, e.g., hypoxia, hypercapnia, heart rate and blood pressure fluctuations, increased sympathetic activity, cortical arousal, or arousal from sleep. These disturbance results in fragmented and non-restorative sleeps. Hence, people with OSA often experience excessive sleepiness, fatigue, poor concentration, irritability, depression, or driving drowsiness, which increase the risk of work/traffic accidents and medical disability. E.g., the odd ratios of traffic accidents for drivers with OSA is from 2.4 to 6.3 times higher than healthy ones. The depression rate of people with OSA is significantly higher than those without OSA (18% vs. 8%), especially on women. OSA is a major global health problem, estimated affecting billions of people (approximately 14% of men and 5% of women) worldwide with significantly increasing trend. However, its heterogenous 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. However, there are wide ranges of learning methods, mostly traditional ML methods such as Random Forest, Support Vector Machine or Gradient Boosting Trees. These techniques also take many different types of data as inputs, e.g., electrocardiography (ECG), oxygen saturation (SpO2), human facial scans, or tracheal respiratory movements. Among them clinical data obtained from Electronic Health Record (EHR) are the most attractive ones due to their ease to collect. However, despite all research efforts, there are still many open questions and potential limitations such as small and biased training/evaluation data, limited clinical features, inconsistent reported results, and lack of DL methods, especially for clinical data.

Our paper aims to two main contributions.

- First, this paper provides a comprehensive review of AI-driven OSA diagnosis, covering different diagnosis problems, input-data types, data biases, pre-processing techniques, and model performance. Our review covers much wider aspects than existing surveys on the field. Thus, it is intended to be useful to other researchers for their future studies.
- Second, we leverage the largest clinical dataset used in OSA prediction to date, approximately 110,000 patients to systematically compare the performance of 39 ML/DL models with16 different feature sets, 4 data imbalance correction techniques and 16 imputation methods. Many of them (especially DL ones) have not been studied before in OSA severity predictions. Our study portrays a broad and detailed view of the performance of OSA predictions from EHR data. The results show that effectively predicting OSA from EHR data is a very challenging problem with particularly low prediction accuracy and biased out- comes towards higher severity levels, while common data preprocessing techniques like imputation, imbalance correction, and feature selection do not really help to boost performance overall. Increasing training data can make it even worse due to many missing data points. 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 specifically designed for OSA.

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

2) What is the evidence you provide to support your claim?

In the review part, we collect a huge set of papers from the literature related to OSA severity predictions. All papers are then carefully reviewed and selected based on multiple criteria such as publication venues, publication years, citations, manually paper selections, and overlapping contents. If there are multiple papers with the same messages, we only select some of them to include in the review to reduce spaces. Overall, there are around 150 papers included into the review. Please note that if a paper does not appear on the review, it does not mean the paper is not good. We may miss the paper. Or there are similar papers but published more recent so we chose the newest ones. All of these papers are grouped into a broad aspect, covering diagnosis problems, types of data and their relevant features, learning techniques, and prediction outcomes. Hence, it provides a wider overview of the field compared to existing review papers. We also cover newest publications until end of 2024. Full detailed on the review can be found in Section 3.

In the comparative study part, we aim at clinical data collected from EHR. In particular, we leverage the largest clinical dataset to date comprising 110,000 patients and 50 selected features for a systematic evaluation of predictive performance across 39 ML/DL models. The analysis encompasses 16 different feature sets, four data imbalance correction techniques, and 16 imputation methods. Notably, several advanced DL models examined here represent novel contributions to OSA severity prediction research, as many of these approaches including most neural network-based models, have not previously been investigated in OSA severity predictions. 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. To the best of our knowledge, our study is the most comprehensive comparative studies in the literature. It is also evaluated on the largest dataset (up to 15 times larger than the second largest dataset), thus reducing possible data biases. Full detailed studies can be found in Section 4.

The findings of this paper have been evaluated/assessed by two medical experts in OSA: Dr. Sebastien Bailly and Prof. Jean-Louis Pepin. Prof. Pepin is regarded as one of the best scientists in Sleep Apnoea worldwide (ranked 3rd by ExpertScape: https://expertscape.com/ex/sleep+apnea).

3) What papers by other authors make the most closely related contributions, and how is your paper related to them?

For the review papers, there exists a few literature survey papers such as [A] and [B]. However, they all focus on one specific aspect. For example, [A] covers all works related to OSA prediction with PSG signals such as ECG or SpO2. [B] only focuses on DL methods for PSG signals. Different to these works, our review takes a broader scope, covering diagnosis problems, types of data and their relevant features, learning techniques, and prediction outcomes. Hence, it provides a wider overview of the field. Moreover, since latest review papers are in 2022, our survey covers newest works until end of 2024.

On the comparative study part, all existing works only cover a very small group of traditional ML techniques. The largest comparative study so far is [C] with around 30 ML algorithms. However, the data used is very small with 1,042 samples and biased toward male patients. DL methods are entirely not studied before by any works. And the effects of different techniques like imputations or data imbalance corrections are completely ignored. Compared to these existing works, 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 apnoea, which can reduce data biases. Moreover, we study a wide range of ML methods including those commonly used and those that have not been studied in OSA research before. Concretely, with 39 different ML and DL models, 17 different sets of features (both automatically selected by different algorithms or manually selected from previous works), 4 data imbalance correction techniques, 16 data imputation methods and 3 prediction problems (binary severity classification, 4-levels severity classification and direct AHI prediction) (i.e., 127,296 combination cases), our study portrays a broad and detailed view of the performance of OSA predictions from EHR data.

References:

[A] Xu, S., Faust, O., Seoni, S., Chakraborty, S., Barua, P.D., Loh, H.W., Elphick, H., Molinari, F., Acharya, U.R.: A review of automated sleep disorder detection. Comput. Biol. Med. (2022)

[B] Mostafa, S.S., Mendon¸ca, F., G. RaveloGarc´ıa, A., Morgado-Dias, F.: A systematic review of detecting sleep apnea using deep learning. Sensors (2019)

[C] Rodrigues Jr, J.F., Pepin, J.-L., Goeuriot, L., Amer-Yahia, S.: An extensive investigation of machine learning techniques for sleep apnea screening. In: CIKM (2020)

4) Have you published parts of your paper before, for instance in a conference?

This paper has not been published before. The submission is entirely original.