Original Article

A machine learning approach for the diagnosis of obstructive sleep apnoea using oximetry, demographic and anthropometric data

Zhou Hao Leong[1], MMed, Shaun Ray Han Loh[1], FAMS, Leong Chai Leow[2], FRACP, Thun How Ong[2], MRCP, Song Tar Toh[1], FAMS

[1]Department of Otorhinolaryngology – Head and Neck Surgery, Singapore General Hospital, [2]Department of Respiratory and Critical Care Medicine, Singapore General Hospital, Singapore

Abstract

Introduction: Obstructive sleep apnoea (OSA) is a serious but underdiagnosed condition. Demand for the gold standard diagnostic polysomnogram (PSG) far exceeds its availability. More efficient diagnostic methods are needed, even in tertiary settings. Machine learning (ML) models have strengths in disease prediction and early diagnosis. We explored the use of ML with oximetry, demographic and anthropometric data to diagnose OSA.

Methods: A total of 2,996 patients were included for modelling and divided into test and training sets. Seven commonly used supervised learning algorithms were trained with the data. Sensitivity (recall), specificity, positive predictive value (PPV) (precision), negative predictive value, area under the receiver operating characteristic curve (AUC) and F1 measure were reported for each model.

Results: In the best performing four-class model (neural network model predicting no, mild, moderate or severe OSA), a prediction of moderate and/or severe disease had a combined PPV of 94%; one out of 335 patients had no OSA and 19 had mild OSA. In the best performing two-class model (logistic regression model predicting no–mild vs. moderate–severe OSA), the PPV for moderate–severe OSA was 92%; two out of 350 patients had no OSA and 26 had mild OSA.

Conclusion: Our study showed that the prediction of moderate–severe OSA in a tertiary setting with an ML approach is a viable option to facilitate early identification of OSA. Prospective studies with home-based oximeters and analysis of other oximetry variables are the next steps towards formal implementation.

Keywords: Diagnosis, machine learning, obstructive sleep apnoea

INTRODUCTION

Obstructive sleep apnoea (OSA) is now well established as a serious condition with significant effects to the patient's health and is also a burden to the public health system and economy. While timely diagnosis and treatment initiation is essential to minimise these deleterious effects, in-laboratory polysomnography (PSG), which is the gold standard for diagnosis of OSA, is resource intensive in both labour and inpatient beds. Appointments for in-laboratory PSG are often insufficient to meet the demand. OSA remains underdiagnosed in many populations, even in developed countries.[1]

In Singapore, the prevalence of OSA is about 30.5%.[2] The average wait for an in-laboratory PSG at the largest tertiary sleep medicine centre in Singapore is 79.8 days, with the average time to treatment initiation approaching 4 months (118.4 days). There are a wide range of other diagnostic methods from the American Association of Sleep Medicine (AASM) level 2 to 4 studies and newer ambulatory sleep study devices. In particular, the WatchPAT ambulatory sleep study has gained

Access this article online

Quick Response Code:

Website:
https://journals.lww.com/SMJ

DOI:
10.4103/singaporemedj.SMJ-2022-170

Correspondence: Dr Zhou Hao Leong,
Associate Consultant, Department of Otorhinolaryngology – Head and Neck
Surgery, Singapore General Hospital, Outram Road, 169608, Singapore.
E-mail: leong.zhou.hao@singhealth.com.sg

Received: 06 Oct 2022 **Accepted:** 25 Jan 2023 **Published:** 02 May 2023

For reprints contact: WKHLRPMedknow_reprints@wolterskluwer.com

How to cite this article: Leong ZH, Loh SRH, Leow LC, Ong TH, Toh ST. A machine learning approach for the diagnosis of obstructive sleep apnoea using oximetry, demographic and anthropometric data. Singapore Med J 2025;66:195-201.

wide acceptance, and adoption of WatchPAT in our centre has led to a reduction in the average time to diagnostic study for appropriate patients to 21 days and the average time to treatment initiation to 46.3 days.[3]

However, these studies have their limitations. Although WatchPAT has been approved for diagnostic use in OSA, a recent validation study by Ioachimescu *et al*.[4] on 500 patients who underwent concurrent PSG and WatchPAT showed high rates of misclassification by WatchPAT. In particular, the diagnostic accuracy for patients deemed to have mild OSA by WatchPAT was only 49.6%, with 30.1% of these patients not having OSA at all and 20.4% of them having moderate to severe OSA. Even for moderate–severe OSA, the positive predictive value (PPV) of WatchPAT was only 76%. As the epidemic of OSA is expected to worsen, the need for a simple but effective diagnostic tool is paramount. Early identification of patients with moderate to severe OSA is essential to timely treatment initiation, as these patients are the most susceptible to subsequent cardiovascular morbidity and mortality.[5]

The use and application of machine learning (ML) algorithms in medicine is rapidly gaining traction. These methods have powerful problem-solving capabilities and the distinct advantage of being able to detect possible interactions among many variables. While there is some overlap between statistical modelling and ML, statistical models are focused on inference of relationships between variables, whereas ML concentrates on making accurate predictions based on available data.[6] As such, there is much interest in the use of ML in the field of clinical prediction and early diagnosis. The application of ML to predict OSA and/or estimate OSA severity is not new. Researchers have taken different approaches, with several exploring the use of demographic data and clinical features such as symptoms and examination findings, achieving accuracy ranging from 55% to 68%.[7-9] Others have used biomedical markers, most commonly oxygen saturation (SpO_2) and electrocardiogram (ECG) signals from wearable devices, with accuracy ranging from 77% to 87%.[10] Combining demographic data and biomarkers is a relatively novel approach, and more studies need to be done to determine its relevance and precision.

Pulse oximetry is a relatively inexpensive investigation modality that is noninvasive and widely available. Overnight oxygen desaturations are characteristic of OSA, and it is thought that oxidative stress due to persistent nocturnal hypoxaemia contributes to the association with cardiovascular morbidity.[11,12] While pulse oximetry is used for the diagnosis of OSA in the paediatric population because of its high PPV, the negative predictive value (NPV) is very low.[13,14] We hypothesise that an ML model trained by combining demographic, anthropomorphic and oximetry data from a population with a high pretest probability of OSA will be able to predict patients with moderate to severe OSA with equal or greater accuracy compared to currently available ambulatory sleep studies. We also aim to evaluate the performance of different ML models in the prediction of OSA.

METHODS

Data and feature selection

This is a retrospective cohort study based on our local Sleep Medicine database. All adult (>18 years) patients who underwent level 1 PSG at a single tertiary sleep medicine centre between January 2017 and December 2020 for suspected OSA and consented to inclusion in our database were included in the study (CIRB Ref No. 2019/2011). Full PSG, demographic and anthropometric data were collected, including age, gender, neck circumference (NC), body mass index (BMI) and Epworth Sleepiness Score (ESS). Oximetry indices collected included 3% oxygen desaturation index (ODI), percentage time of sleep in which the nadir SpO_2 is less than 90% and 85%, and number of desaturations. The data was recorded with the Grael PSG and Grael 4K PSG:EEG Systems (Compumedics, Abbotsford, Victoria, Australia). PSG scoring was performed according to the AASM 2017 guidelines,[15] and the recommended rule was used for the scoring of hypopnoeas. OSA diagnosis and severity was defined by the apnoea–hypopnoea index (AHI), with mild OSA being $5 \leq AHI < 15$, moderate OSA being $15 \leq AHI < 30$ and severe OSA being $AHI \geq 30$.

A total of 3,671 patients were identified. There were 3,068 patients after excluding patients with preexisting comorbidities that may affect oximetry readings, such as heart failure, cardiac arrhythmia, ischaemic heart disease, chronic obstructive pulmonary disease, interstitial lung disease, pulmonary hypertension, epilepsy and parasomnias. Further exclusions were made during data preprocessing and cleaning due to missing data fields and/or uncertain entries, leading to a final data set of 2,996 patients. General patient characteristics, PSG data and results from the bivariate analysis for the study are presented in Table 1. Bivariate analysis was performed with Pearson's rho of AHI versus each individual variable. The final data set was then randomly divided into 80% for the training set (n = 2,397) and 20% for the test set (n = 599). Feature selection was performed using principal component analysis, with seven components selected as accounting for >95% of variance.

Models and algorithms

The algorithms were trained based on three different classification targets. In the first, patients were divided into four groups of none, mild, moderate and severe OSA based on AHI. In the second, patients were divided into two groups of none–mild OSA versus moderate–severe OSA. In the third, patients were divided into two groups of no OSA versus OSA. This would allow us to observe the performance of these models across different clinical contexts. While it would be interesting to see how well the algorithms can discriminate

Table 1. Cohort variables and bivariate analysis.

Variable	*n* (%)/Median±IQR					*P**
	Overall	No OSA	Mild OSA	Moderate OSA	Severe OSA	
No. of patients	2,996	459 (15.3)	692 (23.1)	680 (22.7)	1,165 (38.9)	
Gender						
Male	2,013 (67.2)	253 (12.6)	413 (20.5)	462 (22.9)	885 (43.9)	
Female	983 (32.8)	206 (20.9)	279 (28.4)	218 (22.2)	280 (28.5)	
Oxygen desaturation index	6.5±23.8	0.3±0.8	1.9±3.0	6.7±9.6	33.3±38.5	<0.001
SpO_2 nadir (%)	84.8±14.8	92.8±4.0	88.5±6.2	84.8±9.7	74.4±18.7	<0.001
Time SpO_2 <90% (%)	0.5±4.3	0±0.01	0.1±0.3	0.5±1.7	5.9±16.8	<0.001
Time SpO_2 <85% (%)	0.02±0.8	0±0	0±0.03	0.03±0.3	1.3±5.8	<0.001
Neck circumference (cm)	39±5	36±4	37±5	38±5	41±5	<0.001
Body mass index (kg/m^2)	26.9±7.3	23.8±5.3	25.2±5.7	26.6±6.8	29.5±7.5	<0.001
Epworth Sleepiness Score	9±7	9±7	8±7	9±7	9±7	0.565
Age (yr)	47.0±14.2	39.8±15.0	46.6±14.3	49.2±13.3	48.5±13.8	<0.001

*Bivariate analysis. IQR: interquartile range, OSA: obstructive sleep apnoea

the four classes of OSA, the practical application of such a model would likely be in the quick identification of patients who would benefit most from early treatment initiation (i.e., as a 'rule-in' test). Therefore, we also wanted to observe the performance of these models in discriminating patients with OSA from those without; specifically, patients with moderate–severe OSA should not be missed as there is a high consequence associated with undertreatment.

We employed supervised ML algorithm templates in Orange Data Mining software v3.32.0,[16] including logistic regression (LR), random forest (RF), support vector machine (SVM), adaptive boosting (AdaBoost), gradient boosting (GB), neural network (NN) and k-nearest neighbour (kNN). Models were trained on the training data set and validated with ten-fold cross validation. Trained models were then tested on the test data set, and the classification accuracy, sensitivity (recall), specificity, PPV (precision), NPV, area under the receiver operating characteristic curve (AUROC) and the F1 measure were reported. The F1 measure is the harmonic mean of precision and recall, and therefore takes into account both false positives and false negatives. It is useful, especially in situations where there is an uneven class distribution.

A detailed explanation of the various ML algorithms is beyond the scope of this article. Uddin *et al.*[17] studied and compared the performance of these different supervised ML algorithms. In summary, regression modelling investigates the relationship between the target (dependent) variable and the input (independent) variables. LR is used to find the probability of event = success and event = failure; it is the most commonly used supervised learning model for binary classification and has the benefit of being highly interpretable. RF is a type of decision tree model that uses multiple randomly created decision trees that aggregate a 'vote' for the target outcome. RF can easily handle large volumes of data with numerous variables and has the added benefit of being able to automatically balance data sets when a class is more infrequent than other classes in the data. This is useful in our data set where there is a high prevalence of disease. In our RF models, 100 trees are used with up to five attributes considered at each split. SVM is a commonly used classifier that draws raw data on the *n*-dimensional plane (where *n* = number of variables) and seeks to find the optimum separating hyperplane between the different categories. GB and AdaBoost are algorithms based on multiple decision trees. The GB model is built around sequential decision trees, where the error of the last tree is considered and the decision of every successive tree is built on the mistakes of the previous tree. In our GB model, 1,000 trees are used with a learning rate of 0.1 and the depth of each individual tree is limited to 4. AdaBoost is an ensemble method boosting technique that reassigns weights of each observation based on correct or incorrect predictions in an iterative process, adding new estimators or learners until the model limit is reached. In our AdaBoost model, we used 50 estimators with a learning rate of 1. kNN is a distance-based supervised learning model that locates and identifies the k-number of nearest datapoints to the new unknown datapoint, and predicts a class based on the majority class of the nearest neighbours. In our kNN model, we used a k value of 5. Finally, NNs are a mesh of interconnected nodes, aggregated into layers. Each node is a decision point where a weight is applied to the input information and is passed on to the next if it passes the threshold. The final output then predicts the target class.

RESULTS

No OSA vs. OSA

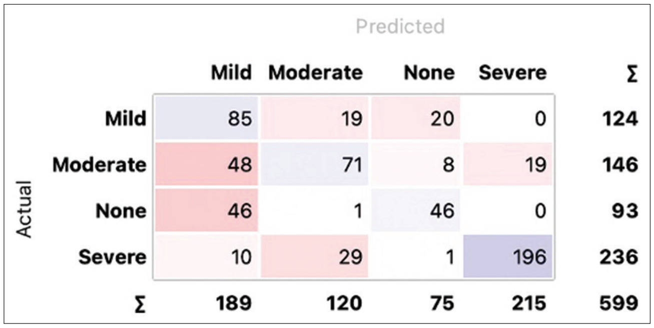
Table 2 summarises the performance of the seven models on the test data set when classifying no OSA versus OSA. The NN model had the best performance with a classification accuracy of 90.2%, AUROC of 0.928 and an F1 value of 0.942. While the sensitivity and PPV for OSA patients were very high at

95.6% and 92.9%, respectively, the specificity and NPV were lower at 60.6% and 72.2%, respectively. This trend was also seen in the rest of the models.

None vs. mild vs. moderate vs. severe OSA

Table 3 summarises the performance of the seven models on the test data set when classifying into the four classes of none, mild, moderate and severe OSA. The overall performance is presented, together with the sensitivity of each specific class prediction. As anticipated in a multi-class model, the classification accuracy was lower across the board, with the best performing model at 66.4%. While the sensitivity of a 'severe OSA' prediction was relatively good at around 80%, performance for the rest of the classes was poorer.

The NN model was the best performing, with an overall classification accuracy of 66.4%. The confusion matrix for test set predictions by the NN model is presented in Figure 1. This illustrates its high sensitivity and PPV in predicting patients with severe OSA. Of 215 patients predicted to have severe OSA, 196 patients did indeed have severe disease. The other 19 patients all had moderate disease. None of them had no or mild disease. The prediction for moderate OSA was also good. Of 120 patients predicted to have moderate OSA, 100 (83.3%) patients did indeed have moderate to severe disease. For the remaining 20 patients, 19 had mild disease and one patient did not have OSA. Taken together, 335 out of 599 patients (55.9%) were predicted to have at least moderate disease with a PPV of 94.0%. Of the 264 patients predicted to have not more than mild disease, 67 (25.4%) were false negative.



		Predicted				
		Mild	Moderate	None	Severe	Σ
Actual	Mild	85	19	20	0	124
	Moderate	48	71	8	19	146
	None	46	1	46	0	93
	Severe	10	29	1	196	236
	Σ	189	120	75	215	599

Figure 1: Confusion matrix for four-class prediction with a neural network model.

Moderate–severe vs. none–mild OSA

We further examined whether the performance of the models in dichotomising between moderate–severe and none–mild patients can be improved in a two-class classifier. However, the performance was about the same. Table 4 summarises the performance of the seven algorithms on the test data set when classifying none–mild OSA versus moderate–severe OSA. The LR model had the best performance with a classification accuracy of 85.3%, AUROC of 0.927 and an F1 value of 0.880. Sensitivity and specificity were high for all models, with the LR model having a sensitivity of 84.3% and specificity of 87.1%. The PPV for LR model was also very high at 92%. The NPV of the models was relatively lower but showed a modest improvement over the four-class classifier models, with the best NPV seen in the LR model at 75.9% (false negative of 22.3%).

DISCUSSION

In this study, we explored the use of ML modelling of data from a single channel (oximetry) together with demographic and anthropometric data to identify patients who would most likely benefit from early treatment initiation in a population with a high pretest probability for OSA (patients seen for suspected OSA at a tertiary sleep medicine centre). Early identification and treatment initiation for OSA has its practical application in many busy hospitals, where demand for diagnostic PSG often far exceeds availability. In some situations, delay in OSA diagnosis and treatment leads to compromise in other aspects of patient care — for example, surgery may be delayed in patients identified by preoperative anaesthesia evaluation to have high risk of OSA but are unable to get a diagnostic test in time.

The performance of any diagnostic test, in particular, PPV and NPV, is influenced by the prevalence of the disease in the population. In our population, the prevalence of OSA was very high, with 84.3% of patients having at least mild OSA. Even though the two-class models (no OSA vs. OSA) had a high PPV, this is partly due to the high prevalence. The lower specificity and NPV means it is not ideal as a rule-out test in a tertiary setting. A positive test only increases the probability of a true positive from a baseline of 84.3% to 92.9% (in the best performing NN and LR models). This increase is modest at best and indicates that the usefulness in this context is limited.

Table 2. Model performance with classification into two classes — no OSA vs. OSA.

Variable	Neural Network	Random Forest	Logistic Regression	Gradient Boosting	kNN	Support Vector Machine	AdaBoost
Accuracy	0.902	0.860	0.891	0.856	0.858	0.885	0.866
AUROC	0.928	0.889	0.926	0.889	0.848	0.917	0.860
F1	0.942	0.917	0.936	0.914	0.917	0.917	0.921
Sensitivity (Recall)	0.956	0.923	0.945	0.909	0.925	0.974	0.927
PPV (Precision)	0.929	0.912	0.928	0.920	0.909	0.898	0.916
Specificity	0.606	0.521	0.606	0.574	0.500	0.404	0.543
NPV	0.722	0.557	0.671	0.540	0.553	0.745	0.580

AdaBoost: adaptive boosting, AUROC: area under the receiver operating characteristic curve, kNN: k-nearest neighbour, NPV: negative predictive value, OSA: obstructive sleep apnoea, PPV: positive predictive value

Table 3. Model performance with classification into four classes — none, mild, moderate vs. severe OSA.

Variable	Neural Network	Random Forest	Logistic Regression	Gradient Boosting	kNN	Support Vector Machine	AdaBoost
Overall							
Accuracy	0.664	0.659	0.653	0.636	0.623	0.569	0.539
AUROC	0.885	0.877	0.882	0.862	0.830	0.809	0.800
F1	0.670	0.663	0.651	0.634	0.628	0.581	0.543
Sensitivity (Recall)	0.664	0.659	0.653	0.636	0.623	0.569	0.539
PPV (Precision)	0.692	0.670	0.658	0.635	0.644	0.649	0.549
No OSA							
Sensitivity	0.495	0.570	0.677	0.570	0.473	0.624	0.516
PPV	0.613	0.596	0.600	0.589	0.571	0.558	0.511
Mild OSA							
Sensitivity	0.685	0.565	0.589	0.540	0.605	0.726	0.444
PPV	0.450	0.461	0.471	0.462	0.417	0.375	0.374
MOD OSA							
Sensitivity	0.486	0.514	0.397	0.418	0.438	0.308	0.356
PPV	0.592	0.564	0.537	0.496	0.520	0.455	0.380
Severe OSA							
Sensitivity	0.831	0.835	0.835	0.847	0.805	0.627	0.712
PPV	0.912	0.876	0.853	0.830	0.868	0.949	0.760

AdaBoost: adaptive boosting, AUROC: area under the receiver operating characteristic curve, kNN: k-nearest neighbour, OSA: obstructive sleep apnoea, PPV: positive predictive value

Table 4. Model performance with classification into two classes — moderate–severe OSA vs. none–mild OSA.

Variable	Neural Network	Random Forest	Logistic Regression	Gradient Boosting	kNN	Support Vector Machine	AdaBoost
Accuracy	0.840	0.831	0.853	0.833	0.801	0.751	0.813
AUROC	0.923	0.912	0.927	0.906	0.892	0.876	0.902
F1	0.869	0.866	0.880	0.866	0.843	0.772	0.851
Sensitivity (Recall)	0.835	0.851	0.843	0.848	0.848	0.660	0.830
PPV (Precision)	0.906	0.881	0.920	0.885	0.885	0.930	0.863
Specificity	0.848	0.797	0.871	0.806	0.742	0.912	0.765
NPV	0.745	0.752	0.759	0.751	0.719	0.604	0.731

AdaBoost: adaptive boosting, AUROC: area under the receiver operating characteristic curve, kNN: k-nearest neighbour, NPV: negative predictive value, OSA: obstructive sleep apnoea, PPV: positive predictive value

Perhaps what is more interesting is trying to identify patients with moderate disease and above. The prevalence of moderate–severe disease in our study population was 61.6%, which places it in the intermediate range of pretest probability. With the best performing four-class model (NN model predicting either no, mild, moderate or severe OSA), a prediction of moderate and/or severe disease had a combined PPV of 94%, with only one out of these 335 patients having no OSA and 19 having mild OSA. In the best performing two-class model (LR model predicting either no–mild or moderate–severe OSA), 350 patients were predicted as having moderate–severe OSA, with 322 of them correctly classified for a PPV of 92.0%. Of the 28 patients incorrectly classified, only two patients had no OSA, while the rest had mild OSA. The increase in the probability of a true positive (of moderate–severe disease) from 61.6% to the region of 92%–94% is sizeable. Even for patients incorrectly classified as moderate–severe, the vast majority at least had mild OSA.

Due to the implications and complications of untreated OSA, it is unacceptable to miss any case of OSA in a tertiary setting. With specificity and NPV values around 70%–80% across most models, it is likely that this approach can only be used as a 'rule-in' test. After initial outpatient review, patients with high suspicion for OSA may undergo an ambulatory pulse oximetry study and routine collection of anthropometric data. This can be done with minimal waiting time due to low cost and high availability of pulse oximeters. Patients predicted to have moderate–severe OSA can then be initiated on positive airway pressure therapy. Those predicted to have no or mild OSA will then undergo a formal in-laboratory PSG. Based on the data from this study, up to 60% of patients who currently undergo in-laboratory PSG may be able to avoid it.

Behar *et al.*[18] described a similar approach of combining oximetry and demographic information, but as a population-based screening tool for OSA. They used a clinical database of 887 individuals from a representative (non-referred) population sample. Correspondingly, the prevalence of OSA was lower in their population at 43.3% (as opposed to 84.3% in this study). Their LR model achieved an accuracy of 86%, with AUROC of 0.94 and an F1 value of 0.84 for the identification of OSA versus no OSA. PPV was 82%, which represents a significant increase from their pretest probability of 43.3%. Although the populations and clinical question are inherently different, the excellent performance of this approach in both Behar *et al*.'s study and this study shows that combining oximetry and demographic data has much promise, both in the context of population-based screening and in the early identification of moderate–severe disease in a tertiary care setting.

Besides its immediate clinical relevance, the strengths of this study also include our large study population. Data set size is critical in determining the performance of an ML model. Typically, large data sets lead to better classification, as there is a risk of overfitting in smaller data sets.[19] Our final data set size of 2,996 patients represents one of the largest used for the application of ML in OSA. As all patients undergoing evaluation at our centre are routinely recruited into our database, the data set is also highly representative of our target population.

In addition, we included a wide range of commonly used supervised ML algorithms. Not surprisingly, the NN models generally performed better than the rest of the models with only some exceptions. The nature of NN makes it suitable for the analysis of large data sets. However, the drawback of the so-called 'flexible' ML algorithms, such as NN, RF and Boosting, is their reduced interpretability, or the so-called 'Black box' dilemma of artificial intelligence.[20] Physicians and patients alike may be hesitant to trust and adopt something that they do not fully understand. To that end, the use of more restrictive but interpretable algorithms such as SVM, kNN and regression models may be more acceptable. Although these models have slightly lower accuracy in general, we have demonstrated that their performance is comparable when there is a large volume of data.

The limitations of this study include its retrospective nature and the derivation of oximetry data from an in-laboratory PSG instead of a home oximeter. Home oximeters may underestimate the true extent of desaturations due to their inability to correct for sleep duration and their reduced sensitivity to desaturations compared to PSG-derived oximetry. If this approach is to be implemented for use in clinical practice, prospective studies utilising data from home oximeters should be done. The data from the home-based oximeter model prediction should then be validated against the gold standard in-laboratory PSG.

In addition, desaturation indices such as ODI, SpO_2 nadir and cumulative time under a certain saturation level are not the only variables that can be extracted from oximetry. There are various technical approaches to analysing oximetry that may better capture the physiological complexity in OSA. These include characteristics of desaturations (e.g., desaturation depth, duration and area), time series statistics (mean, variance, skew, kurtosis, minimum and cumulative times), analysis of power spectral distributions (i.e., frequency domain analysis), and nonlinear analysis (e.g., regularity and complexity). Terrill[21] outlined these approaches in his enlightening review and postulated that information from these different domains may be complementary to just desaturations alone. Including variables from all four domains may, therefore, contribute to improving model accuracy and performance. In addition, other variables such as pulse rate and pulse rate variability can also be extracted from pulse oximeter readings. This abundance of data points provides a rich mining field for ML algorithms to further refine their performance.

In conclusion, our study shows that combining oximetry, demographic and anthropomorphic data for the prediction of moderate–severe OSA in a tertiary sleep clinic population with an ML approach is a viable option to facilitate early identification and treatment initiation in these patients. Both the four-class and two-class classifiers predict moderate–severe OSA with a PPV of 92%–94%, with most of the wrongly classified cases having mild OSA. It has excellent potential utility as a 'rule-in' test and can possibly reduce the need for diagnostic PSG by 60%. Prospective studies utilising a home-based oximeter and analysis of other pulse oximeter variables are the next steps towards formal implementation of such an approach.

Financial support and sponsorship

Nil.

Conflicts of interest

Toh ST is a member of the *SMJ* Editorial Board, and was thus not involved in the peer review and publication decisions of this article.

REFERENCES

1. Benjafield AV, Ayas NT, Eastwood PR, Heinzer R, Ip MSM, Morrell MJ, *et al*. Estimation of the global prevalence and burden of obstructive sleep apnoea: A literature-based analysis. Lancet Respir Med 2019;7:687-98.
2. Tan A, Cheung YY, Yin J, Lim WY, Tan LW, Lee CH. Prevalence of sleep-disordered breathing in a multiethnic Asian population in Singapore: A community-based study. Respirology 2016;21:943-50.
3. Phua CQ, Jang IJ, Tan KB, Hao Y, Senin SRB, Song PR, *et al*. Reducing cost and time to diagnosis and treatment of obstructive sleep apnea using ambulatory sleep study: A Singapore sleep centre experience. Sleep Breath 2021;25:281-8.
4. Ioachimescu OC, Allam JS, Samarghandi A, Anand N, Fields BG, Dholakia SA, *et al*. Performance of peripheral arterial tonometry–based testing for the diagnosis of obstructive sleep apnea in a large sleep clinic cohort. J Clin Sleep Med 2020;16:1663-74.
5. Yeghiazarians Y, Jneid H, Tietjens JR, Redline S, Brown DL, El-Sherif N, *et al*. Obstructive sleep apnea and cardiovascular disease: A scientific statement from the American Heart Association. Circulation 2021;144:E56-67.

6. Bzdok D, Altman N, Krzywinski M. Points of significance: Statistics versus machine learning. Nat Methods 2018;15:233-4.
7. Holfinger SJ, Lyons MM, Keenan BT, Mazzotti DR, Mindel J, Maislin G, *et al*. Diagnostic performance of machine learning-derived OSA prediction tools in large clinical and community-based samples. Chest 2022;161:807-17.
8. Bozkurt S, Bostanci A, Turhan M. Can statistical machine learning algorithms help for classification of obstructive sleep apnea severity to optimal utilization of polysomnography resources? Methods Inf Med 2017;56:308-18.
9. Ramesh J, Keeran N, Sagahyroon A, Aloul F. Towards validating the effectiveness of obstructive sleep apnea classification from electronic health records using machine learning. Healthcare (Basel) 2021;9:1450.
10. Ramachandran A, Karuppiah A. A survey on recent advances in machine learning based sleep apnea detection systems. Healthcare (Switzerland) 2021;9:914.
11. Dyugovskaya L, Lavie P, Lavie L. Increased adhesion molecules expression and production of reactive oxygen species in leukocytes of sleep apnea patients. Am J Respir Crit Care Med 2002;165:934-9.
12. Lavie L. Obstructive sleep apnoea syndrome – An oxidative stress disorder. Sleep Med Rev 2003;7:35-51.
13. Brouillette RT, Morielli A, Leimanis A, Waters KA, Luciano R, Ducharme FM. Nocturnal pulse oximetry as an abbreviated testing modality for pediatric obstructive sleep apnea. Pediatrics 2000;105:405-12.
14. Nixon GM, Kermack AS, Davis GM, Manoukian JJ, Brown KA, Brouillette RT. Planning adenotonsillectomy in children with obstructive sleep apnea: The role of overnight oximetry. Pediatrics 2004;113:e19-25.
15. Berry RB, Brooks R, Gamaldo C, Harding SM, Lloyd RM, Quan SF, *et al*. AASM scoring manual updates for 2017 (version 2.4). J Clin Sleep Med 2017;13:665-6.
16. Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, *et al*. Orange: Data Mining Toolbox in Python. J Mach Learn Res 2013;14:2349-53.
17. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 2019;19:281.
18. Behar JA, Palmius N, Li Q, Garbuio S, Rizzatti FPG, Bittencourt L, *et al*. Feasibility of single channel oximetry for mass screening of obstructive sleep apnea. EClinicalMedicine 2019;11:81-8.
19. Althnian A, AlSaeed D, Al-Baity H, Samha A, Dris AB, Alzakari N, *et al*. Impact of dataset size on classification performance: An empirical evaluation in the medical domain. Appl Sci 2021;11:796.
20. Poon AIF, Sung JJY. Opening the black box of AI-medicine. J Gastroenterol Hepatol 2021;36:581-4.
21. Terrill PI. A review of approaches for analysing obstructive sleep apnoea-related patterns in pulse oximetry data. Respirology 2020;25:475-85.