Hidden Hazards Beneath Cross-Validation Methods in Machine Learning-Based Sleep Apnea Detection

Daniele Padovano[1], Arturo Martinez-Rodrigo[1], Jose M Pastor[1], Jose J Rieta[2], Raul Alcaraz[1]

[1] Research Group in Electronic, Biomedical and Telecomm. Eng., Univ. of Castilla-La Mancha, Spain
[2] BioMIT.org, Electronic Engineering Department, Universitat Politecnica de Valencia, Spain

Abstract

Obstructive sleep apnea (OSA) is a respiratory disorder highly correlated with multiple cardiovascular diseases. In the last two decades, several alternatives have been proposed to palliate the limitations of polysomnography, the current gold standard for OSA diagnosis. Such alternatives were mainly based on the heart rate variability in combination with machine learning (ML) techniques, obtaining promising results. However, the majority of these works used a cross-validation approach for the validation of the proposed methods, and rarely tested them on external sources of newly added data. Hence, some of the most common algorithms found in the state of the art have been evaluated with cross-validation and external validation in this work. The obtained results have raised important concerns on the real performance shown by the typical ML-based OSA detection methods in more realistic scenarios.

1. Introduction

Obstructive sleep apnea (OSA) is a condition characterized by multiple respiratory arrests during sleep [1], which prevalence is considered high, affecting from 9 to 38% of the general population [1]. Patients suffering from OSA can describe feelings of grogginess and daytime sleepiness, which in turn can provoke bad job or school performance, family problems, and road accidents in the most severe cases [1]. Besides, this syndrome is also correlated with multiple cardiovascular diseases (CVD), such as atrial fibrillation, strokes, coronary diseases, and so forth [2]. Since CVD are the leading cause of global dead every year [3], the early detection of comorbilities like OSA have gained increasing popularity in recent years [4].

The OSA syndrome is highly infra-diagnosed, which may be in part because polysomnography (PSG) is still considered the gold standard method for OSA detection [1]. PSG is a resource-intensive procedure that requires access to specialist facilities, like sleep laboratories, as well as the presence of a clinical expert to monitor the patient's sleep overnight. On the grounds of the elevated cost and complexity associated to PSG, the application of this procedure is limited to the most wealthy population, aggravating the misdiagnosis problem. For this reason, several alternatives have been proposed in the last two decades [4]. Such alternatives were mainly based on single-sensor approaches, being the single-lead ECG the most effective according to the latest reviews on OSA detection [5]. In this regard, the heart rate variability (HRV) has been the most frequently employed physiological measurement to detect OSA episodes, due to its clinical relationship with the autonomic nervous system, which is in charge of breathing control [6].

Nevertheless, despite the rich amount of approaches published in the state of the art, no one has been actually considered for clinical practice purposes to finally replace the gold standard PSG. In this regard, the latest research on OSA detection has been fundamentally conducted through machine learning-based classifiers [4], most of them following a conspicuous, consolidated pattern, i.e., signal processing, feature extraction, classification and model validation [5]. In such pattern, the validation stage is almost exclusively carried out with cross-validation methods, but seldom with external sources of newly added data [7]. Hence, the main goal of the present work is to evaluate the most typical machine learning (ML) approaches for OSA detection under a more realistic validation scenario, closer to clinical practice circumstances.

2. Methodology

2.1. Databases

Three freely available databases from the PhysioNet's repository were employed, i.e., the CinC Challenge 2000 (Apnea-ECG) [8], the MIT Polysomnographic Database (MIT-BIH) [9], and the St.Vicent's University Hospital/University College of Dublin (UCD-DB) [10], all containing several ECG recordings with apneic annotations.

The Apnea-ECG database consists of 70 ECG recordings, 7 to 9 hours length, coming from 30 male and 5 fe-

male subjects within 27 to 63 years old. This database includes annotations made by clinical experts in a minute-by-minute basis, assessing whether right at the beginning of a minute of the ECG recording there was an apneic epoch (A-labeled), or a normal epoch (N-labeled).

The MIT-BIH database consists of 18 PSG recordings between 2 and 7 hours of duration obtained from 16 male subjects within 32 and 56 years of age. Annotations were also provided by clinical experts under a similar criteria to Apnea-ECG's database, but every 30 seconds.

Eventually, the UCD-DB consists of 25 full overnight PSG recordings coming from 21 male and 4 female subjects within 28 and 68 years of age. The single-lead ECG was annotated in real-time following the Rechtschaffen and Kales rules, indicating cardiorespiratory events of varying kinds apart from apnea epochs.

In view of the differences between annotation systems, these were adapted to the most limiting in time resolution, i.e., Apnea-ECG. Specifically, the MIT-BIH was re-labeled by retrieving the original annotation every two blocks of 30s, whereas the UCB-DB was re-labeled by retrieving the original annotation in a minute-by-minute basis. Thus, all databases were coherently annotated under the same labeling criteria.

2.2. Signal processing

The ECG recordings were firstly re-sampled at 500Hz. Secondly, the baseline wander and high frequency noise were substracted from the original signal through a band-pass, second-order Chebyshev filter with cut-off frequencies of 0.5 Hz and 100 Hz. In the third place, the R-peak detection was performed using the Pan-Tompkins algorithm [11]. Eventually, the processed ECG recordings were subdivided in segments of one-minute length, and then submitted to a thorough visual revision, discarding those segments that presented excessive noise or artifacts.

2.3. Feature extraction

From each ECG segment, multiple features were extracted following the guidelines found in the state of the art. Most of these were already contemplated by the Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology (Task Force) [6]. Such features can be approached from three different perspectives, i.e., the time domain, the frequency domain, and the complexity domain. In the time domain, the most commonly extracted features were the mean value of the HRV (MEAN), the standard deviation (SDSD), the median value (MED), and many others (see Table 1); whereas in the frequency domain, the Task Force contemplated three different bands of power spectral density (PSD), these were the very low frequency band (VLF, 0.003 – 0.04 Hz), the low frequency band (LF, 0.04 - 0.15 Hz), and the high frequency band (HF, 0.12 - 0.4 Hz). Such bands were computed with both fast Fourier transforms (FFT) and the Lomb-Scargle (LS) periodgram [12]. Last but not least, it was found that the sample entropy (SampEn), and its bidimensional form, the quadratic sample entropy (QSampEn), were two of the most useful features to detect apneic episodes [13]. However, further alternative forms of entropy have gained some popularity in recent years, such as the dispersion entropy (DispEn), the distribution entropy (DistEn), and the fuzzy entropy (FuzzEn) [14], among others (see Table 1). In addition, some features of the recurrence plot of HRV have also been extracted, such as the recurrence rate (R), the divergence (DIV), and the determinism (DET) [15], among others.

Domain	Feature	Description
Time	MAX	Maximum of RRi
	MIN	Minimum of RRi
	MEAN	Mean of RRi
	MED	Median of RRi
	SDNN	Standard deviation of normal RRi
	SDSD	Standard deviation of the differences between adjacent RRi
	RMSSD	Root mean square of differences between adjacent RRi
	NN50	Pairs of adjacent NNi differing by more than 50 ms
	pNN50	Determined by dividing NN50 by the total number of all NNi
	IQR	Interquartile range
Frequency	VLF	FFT very low frequency component
	LF	FFT low frequency component
	HF	FFT high frequency component
	LS-VLF	LSP very low frequency component
	LS-LF	LSP low frequency component
	LS-HF	LSP high frequency component
Complexity	SampEn	Sample entropy
	QSampEn	Quadratic SampEn
	NPSampEn	Non-parametric SampEn
	DispEn	Dispersion Entropy
	DistEn	Distribution Entropy
	FuzzEn	Fuzzy Entropy
	MFuzzEn	Measure of Fuzzy Entropy
	REC	RP Recurrence rate
	DET	RP Determinism
	ENTR	RP Shannon entropy
	L	RP average diagonal line length
	DIV	RP divergence

Table 1. Summary of features extracted from HRV.

2.4. Classification tools

Five different machine learning classifiers were implemented in this work, following the configuration employed

in the majority of works present in the literature. Namely, decision tree (DT), support vectors machine (SVM), the k-nearest neighbors (KNN) algorithm, and two random forest-based classifiers were implemented, i.e., the adaptive boosting (ADA), and bootstrap aggregation (BAG).

2.5. Validation tools

The generated models were validated through the typical performance parameters of accuracy (Ac), sensitivity (Se), and specificity (Sp) by following two different frameworks, i.e., cross-validation [16] and external validation [17]. First, the 10-fold cross-validation was employed to reproduce the methods and results published in the state of the art. Secondly, the same models were tested with databases totally alien to the original training set. This was conducted in 6 different experiments, corresponding to all possible combinations of databases. More specifically, in all experiments, a model trained with a certain database was validated with the remaining ones.

3. Results

The obtained results are presented in Tables 2- 6, where the testing set is deduced as the complementary combination of the training set. As can be observed, external validation results were drastically lower than those obtained with the typical 10-fold cross-validation method. The overall difference between validation methods was nearly a 20% in Ac. However, in the worst case, almost a 40% difference between validation methods was present when the SVM model was trained with the UCD-DB and validated with the Apnea-ECG & MIT-BIH databases.

Training dataset	Cross-validation (k=10)			External validation		
	Ac (%)	Se (%)	Sp (%)	Ac (%)	Se (%)	Sp (%)
Apnea-ECG	75.95	75.42	76.48	62.46	63.72	57.09
MIT-BIH	73.51	72.29	74.73	59.41	66.19	44.02
UCD-DB	74.75	76.09	73.41	52.11	42.87	66.64
MIT-BIH & UCD-DB	69.72	69.54	69.90	57.15	57.88	55.98
Apnea-ECG & UCD-DB	74.36	74.31	74.42	56.45	54.20	59.85
Apnea-ECG & MIT-BIH	74.66	74.28	75.04	62.38	63.58	52.38

Table 2. Results obtained with DT.

Training dataset	Cross-validation (k=10)			External validation		
	Ac (%)	Se (%)	Sp (%)	Ac (%)	Se (%)	Sp (%)
Apnea-ECG	78.56	81.23	75.90	72.64	81.67	34.21
MIT-BIH	73.48	75.75	71.20	69.85	94.38	14.21
UCD-DB	78.12	75.82	80.43	41.41	12.32	87.13
MIT-BIH & UCD-DB	72.85	72.78	72.92	66.55	84.10	38.79
Apnea-ECG & UCD-DB	77.47	79.56	75.37	59.69	74.16	37.81
Apnea-ECG & MIT-BIH	77.72	77.66	77.77	77.52	83.20	30.16

Table 3. Results obtained with SVM.

Training dataset	Cross-validation (k=10)			External validation		
	Ac (%)	Se (%)	Sp (%)	Ac (%)	Se (%)	Sp (%)
Apnea-ECG	81.73	79.68	83.78	64.19	64.26	63.87
MIT-BIH	79.43	79.42	79.42	64.31	73.87	42.63
UCD-DB	79.12	78.57	79.66	49.36	31.08	78.11
MIT-BIH & UCD-DB	74.97	75.00	74.94	61.83	62.94	60.06
Apnea-ECG & UCD-DB	80.23	80.80	79.66	62.40	57.77	69.41
Apnea-ECG & MIT-BIH	80.38	79.67	81.09	65.15	66.51	53.77

Table 4. Results obtained with KNN.

Training dataset	Cross-validation (k=10)			External validation		
	Ac (%)	Se (%)	Sp (%)	Ac (%)	Se (%)	Sp (%)
Apnea-ECG	81.37	81.42	81.32	68.11	69.96	60.24
MIT-BIH	78.46	78.43	78.49	64.22	75.09	39.57
UCD-DB	80.16	80.54	79.79	50.45	28.80	84.49
MIT-BIH & UCD-DB	75.21	74.66	75.76	65.42	66.74	63.33
Apnea-ECG & UCD-DB	79.74	79.83	79.64	64.66	59.23	72.88
Apnea-ECG & MIT-BIH	80.66	80.72	80.59	71.21	73.89	48.91

Table 5. Results obtained with ADA.

Training dataset	Cross-validation (k=10)			External validation		
	Ac (%)	Se (%)	Sp (%)	Ac (%)	Se (%)	Sp (%)
Apnea-ECG	83.42	82.16	84.67	67.20	69.05	59.32
MIT-BIH	81.16	80.59	81.74	63.90	74.62	39.58
UCD-DB	82.24	82.16	82.32	50.80	28.73	85.51
MIT-BIH & UCD-DB	77.31	78.08	76.53	63.33	64.26	61.86
Apnea-ECG & UCD-DB	81.96	81.54	82.39	64.34	60.47	70.21
Apnea-ECG & MIT-BIH	82.40	81.19	83.61	68.76	70.65	53.08

Table 6. Results obtained with BAG.

4. Discussion

In view of the obtained results, regardless of the employed classifier, models validated with cross-validation presented more optimistic results compared to those obtained under the external validation approach. This means that the same models trained and validated with recordings coming from similar subjects do not properly generalize the OSA detection problem in newly added data, such as in the clinical practice scenario. The bias present in cross-validation methods was strongly influential in the numerical results, giving misleading insights of the real generalization capability of the assessed models. Thus, it is strongly recommended to contrast validation results with further external sources of information whenever possible.

Although the external validation method may bring lower results, these are more realistic and faithful. In fact, this latter approach is advocated by Transparent Reporting of a multivariate prediction model for Individual Prognosis Or Diagnosis (TRIPOD) initiative [17]. Hence OSA detectors based on solely one of the databases involved in this paper are susceptible to bias and lack of enough generalization in clinical practice, which is in line with former studies on the well-known Apnea-ECG database [18].

On the other hand, among the variations in the performance parameters, it is possible to observe that the smallest databases provided also the lowest results in both cross-validation and external validation approaches, thus indicat-

ing a clear correlation between generalization capability and model performance. The UCD-DB database meant the smallest training dataset in terms of hours of ECG recordings, whereas the Apnea-ECG was the biggest one and the models trained on it provided higher results and more stable values of Se and Sp. Hence, for proper generalization of ML-based OSA detectors, further and bigger public databases are still required.

5. Conclusions

The present work has evaluated the most common ML-based methods in the context of OSA detection from the single-lead ECG. Multiple features from the HRV have been extracted in accordance to the guidelines found in the state of the art. The generated models have been validated from two different perspectives, the popular 10-fold cross-validation and the external validation approach. The results have proven that the models based on cross-validation were not sufficiently general to properly discern between apnea and normal epochs in newly added data and external validation is essential to provide a realistic view of their performance. Moreover, larger public databases seem still to be required to improve the generalization ability of ML-based OSA detectors.

Acknowledgments

This research has received financial support from public grants PID2021-00X128525-IV0 and PID2021-123804OB-I00 of the Spanish Government 10.13039/501100011033 jointly with the European Regional Development Fund (EU), SBPLY/17/180501/000411 and SBPLY/21/180501/000186 from Junta de Comunidades de Castilla-La Mancha, and AICO/2021/286 from Generalitat Valenciana. Moreover, Daniele Padovano holds a predoctoral scholarship 2022-PRED-20642, which is co-financed by the operating program of European Social Fund (ESF) 2014-2020 of Castilla-La Mancha.

References

[1] Benjafield AV, Ayas NT, Eastwood PR, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. The Lancet Respiratory Medicine 2019;7(8):687–698.

[2] Zapater A, Sánchez-de-la Torre M, Benítez ID, et al. The effect of sleep apnea on cardiovascular events in different acute coronary syndrome phenotypes. American Journal of Respiratory and Critical Care Medicine Jul 2020;.

[3] Organization WH, et al. Noncommunicable diseases country profiles 2018. Publications on NCDs 2018;.

[4] JeyaJothi ES, Anitha J, Rani S, Tiwari B. A comprehensive review: Computational models for obstructive sleep apnea detection in biomedical applications. BioMed Research International 2022;2022:e7242667.

[5] Bahrami M, Forouzanfar M. Sleep apnea detection from single-lead ECG: A comprehensive analysis of machine learning and deep learning algorithms. IEEE Transactions on Instrumentation and Measurement 2022;1–1.

[6] Camm J. Task force of the european society of cardiology and the north american society of pacing and electrophysiology. Heart rate variability: standarts of measurement, physiological interpretation and clinical use. Circulation 1996;93:1043–1065.

[7] Varon C, Caicedo A, Testelmans D, Buyse B, Huffel SV. A novel algorithm for the automatic detection of sleep apnea from single-lead ECG. IEEE Transactions on Biomedical Engineering September 2015;62(9):2269–2278.

[8] Penzel T, Moody GB, Mark RG, Goldberger AL, Peter JH. Apnea-ECG database, 2000. URL `https://physionet.org/content/apnea-ecg/`.

[9] Ichimaru Y, Moody GB. MIT-BIH polysomnographic database, 1992. URL `https://physionet.org/content/slpdb/`.

[10] McNicholas W, Doherty L, Ryan S, Garvey J, Boyle P, Chua E. St. Vincent's University Hospital / University College Dublin Sleep Apnea Database, 2004. URL `https://physionet.org/content/ucddb/`.

[11] Pan J, Tompkins WJ. A real-time QRS detection algorithm. IEEE Transactions on Biomedical Engineering March 1985;BME-32(3):230–236.

[12] Scargle JD. Studies in astronomical time series analysis. II-Statistical aspects of spectral analysis of unevenly spaced data. The Astrophysical Journal 1982;263:835–853.

[13] Liang D, Wu S, Tang L, Feng K, Liu G. Short-term HRV analysis using nonparametric sample entropy for obstructive sleep apnea. Entropy 2021;23(3):267.

[14] Li P, Liu C, Li K, et al. Assessing the complexity of short-term heartbeat interval series by distribution entropy. Medical Biological Engineering Computing 2015;53(1):77–87.

[15] Marwan N, Carmen Romano M, Thiel M, Kurths J. Recurrence plots for the analysis of complex systems. Physics Reports January 2007;438(5):237–329.

[16] Browne MW. Cross-validation methods. Journal of Mathematical Psychology March 2000;44(1):108–132.

[17] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. British Journal of Surgery 2015;102:148–158.

[18] Papini GB, Fonseca P, Margarito J, et al. On the generalizability of ECG-based obstructive sleep apnea monitoring: merits and limitations of the Apnea-ECG database. In 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2018; 6022–6025.

Address for correspondence:

Daniele Padovano
Campus Universitario S/N. 16071 – Cuenca (Spain)
daniele.padovano@uclm.es