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

Daniele Padovano, Arturo Martinez-Rodrigo, José M Pastor, José J Rieta, Raúl Alcaraz

Research Group in Electronic, Biomedical and Telecommunication Engineering, University of Castilla-La Mancha, Spain

Background and Aim. Obstructive sleep apnea (OSA) is a respiratory disorder highly correlated with severe cardiovascular diseases. Although polysomnography is still considered the gold standard for OSA detection, its expensive requirements raised the need for alternative methods. In this regard, heart rate variability (HRV) and machine learning (ML) have gained popularity in recent years. Consequently, several works employed publicly available databases to train their models in a reproducible way. However, most of them also relied their validation on cross-validation methods using solely a single database. This hitherto fact brought forward the aim of the present work, i.e., how these models would have performed if they were tested in more realistic scenarios, such in an alien database, or even in the clinical practice.

Methods. The Apnea-ECG, MIT-BIH Polysomnographic, and the University College Dublin databases were re-annotated under the same labeling criteria. The corresponding ECG recordings were segmented into one-minute length epochs to extract the HRV, as well as its most representative features according to the state of the art. Then, various well-known ML classifiers were trained with different combinations of balanced subsets, computing the 10-fold cross-validation for each model. Eventually, these models were also tested on the remaining datasets, which were alien to the original training sets.

Results. External validation results have shown 10-40% lower performance than 10-fold cross-validation regardless of the selected model (see table).

Conclusions. The obtained results suggest the need for larger datasets to properly generalize the apnea detection problem, especially in those ML models trained and tested with cross-validation on a single database, which are suspected to be over-fitted.

Model/Performance	10-fold cross-validation			External validation		
	Ac (%)	Se (%)	Sp (%)	Ac (%)	Se (%)	Sp (%)
Support-Vector Machine	78.12	75.82	80.43	41.41	12.32	87.13
K-Nearest Neighbours	79.12	78.57	79.66	49.36	31.08	78.11
ADA Boost Ensemble	80.16	80.54	79.79	50.45	28.80	84.49