KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Deviaene M., Testelmans D., Borzee P., Buyse B., Van Huffel S., Varon C.,. **Feature Selection Algorithm based on Random Forest applied to Sleep Apnea Detection** , in Proc. of the 41th Annual International Conference of the Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, Aug. 2019, pp. 1-4
Archived version	Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher
Published version	/
Journal homepage	https://embc.embs.org/2019/
Author contact	your email margot.deviaene@esat.kuleuven.be your phone number *+ 32 (0)16 32 89 60*
Abstract	This paper presents a new feature selection method based on the changes in out-of-bag (OOB) Cohen kappa values of a random forest (RF) classifier, which was tested on the automatic detection of sleep apnea based on the oxygen saturation signal (SpO2). The feature selection method is based on the RF predictor importance defined as the increase in error when features are permuted. This method is improved by changing the classification error into the Cohen kappa value, by adding an extra factor to avoid correlated features and by adapting the OOB sample selection to obtain a patient independent validation. When applying the method for sleep apnea classification, an optimal feature set of 3 parameters was selected out of 286. This was half of the 6 features that were obtained in our previous study. This feature reduction resulted in an improved interpretability of our model, but also a slight decrease in performance,

	without affecting the clinical screening performance. Feature selection is an important issue in machine learning and especially biomedical informatics. This new feature selection method introduces interesting improvements of RF feature selection methods, which can lead to a reduced feature set and an improved classifier interpretability.
IR	/

(article begins on next page)

Feature Selection Algorithm based on Random Forest applied to Sleep Apnea Detection

Margot Deviaene[1], Dries Testelmans[2], Pascal Borzée[2], Bertien Buyse[2], Sabine Van Huffel[1] and Carolina Varon[1]

***Abstract*— This paper presents a new feature selection method based on the changes in out-of-bag (OOB) Cohen kappa values of a random forest (RF) classifier, which was tested on the automatic detection of sleep apnea based on the oxygen saturation signal (SpO_2). The feature selection method is based on the RF predictor importance defined as the increase in error when features are permuted. This method is improved by changing the classification error into the Cohen kappa value, by adding an extra factor to avoid correlated features and by adapting the OOB sample selection to obtain a patient independent validation. When applying the method for sleep apnea classification, an optimal feature set of 3 parameters was selected out of 286. This was half of the 6 features that were obtained in our previous study. This feature reduction resulted in an improved interpretability of our model, but also a slight decrease in performance, without affecting the clinical screening performance. Feature selection is an important issue in machine learning and especially biomedical informatics. This new feature selection method introduces interesting improvements of RF feature selection methods, which can lead to a reduced feature set and an improved classifier interpretability.**

I. Introduction

One of the most important problems in machine learning applications is to decide which features to use to build your model. Often, a large number of relevant features can be defined which might have an added value to the model. Some of these features could, however, be less informative than expected or highly correlated. Therefore, feature selection methods are used to select an optimal feature set which contains the most relevant features, while avoiding redundancy. This can avoid overfitting to the training set and can lead to an improved performance. Moreover, a reduced feature set leads to a faster feature extraction and model optimization and will improve the interpretability of the model [1]. This is very important when developing clinical screening algorithms for wearable devices, since the computational cost should be as low as possible and clinicians are interested in the underlying physiological mechanisms of the machine learning model.

Multiple studies have been focussed on the development of feature selection methods. Simple ranking of features can be performed using statistical measures between the feature values and the output variable, such as the correlation value or the F-score for classification. This ranking, however, does not take into account the correlation between features. More elaborate feature set selection methods are needed for this, such as the minimal redundancy maximal relevance (mRMR) method [2] or a forward or backward wrapper [1]. The mRMR algorithm does not provide a clear determination of the optimal amount of features, and can thus still retain redundant features. The wrapper methods on the other hand, are computationally very expensive when there is a large number of features. Random forest (RF) classifiers, however, provide a faster feature selection method based on the feature importance, which can be measured by calculating the increase in misclassification rate when the out-of-bag (OOB) values of features are permuted. Therefore, the model should not be recomputed for all considered feature sets.

This paper proposes further improvements to this RF feature selection method based on the decrease of permuted OOB Cohen kappa coefficients and the correlation between features in order to further prune feature sets, while allowing a slight decrease in performance. The method will be tested on the detection of sleep apnea.

Sleep apnea hypopnea syndrome (SAHS) is the most common sleep related breathing disorder, which causes cessations of breathing during sleep. Nevertheless, this disorder often remains undiagnosed due to the costly and cumbersome diagnosis consisting of a full night assessment in a hospital during which several physiological signals are measured, called a polysomnography (PSG). This PSG is afterwards manually scored by a sleep technologist and the severity of sleep apnea is then expressed as the apnea-hypopnea index (AHI), which is a count of the number of respiratory events per hour of sleep [3]. Patients suffering from sleep apnea should be diagnosed and treated as early as possible, since studies have shown that untreated sleep apnea can increase the risk of developing cardiovascular diseases [4].

Therefore, research has focused on developing automated detection systems that could be used in a home environment for the screening of SAHS. Easy to measure physiological signals have been considered, such as the electrocardiogram

*This work was supported by: Agentschap Innoveren en Ondernemen (VLAIO) Project #: SWT 150466 - OSA+. imec funds 2017. imec ICON projects: ICON HBC.2016.0167. European Research Council: The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC Advanced Grant: BIOTENSORS (n 339804). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information. Carolina Varon is a postdoctoral fellow of the Research Foundation-Flanders (FWO).

[1]Margot Deviaene, Sabine Van Huffel and Carolina Varon are with the Department of Electrical Engineering-ESAT, STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, and imec, KU Leuven, B-3001 Leuven, Belgium, margot.deviaene@esat.kuleuven.be

[2]Dries Testelmans, Pascal Borzée and Bertien Buyse are with the UZ Leuven, Department of Pneumology , Leuven, Belgium.

(ECG), respiration, oxygen saturation (SpO_2) [5] and pulse photoplethysmography (PPG) [6], or a combination of these signals [7].

The aim of this study is to improve the selection of the optimal feature set for SpO_2 based sleep apnea detection in [5] by implementing the proposed feature selection algorithm.

II. Methods: Feature selection algorithm

This section gives an overview of the proposed RF feature selection method. First a general overview of RF and its standard feature selection method is given. Then the proposed modifications to this algorithm are discussed; starting with the use of the Cohen kappa value as a performance estimate, then the inclusion of feature correlations and the use of patient independent validation.

A. Random forest and feature selection

A random forest classifier is an ensemble of decision trees in which each tree is trained on a random subset of data points (bagging) and a random sampling of features is performed at each node of the tree [8]. The samples not used for training per tree are called out-of-bag samples and will be used to assess the performance of the classifier. For the implementation of this study the `Treebagger` class in `Matlab` was used. All RFs were trained with 100 trees.

As proposed in [8], feature selection in a RF classifier can be performed by building a backwards wrapper in which for each run the feature with the lowest predictor importance (PI) is removed. This PI is calculated for each feature as the increase in misclassification error when the values of the feature under investigation were randomly permuted. When the feature has a high relevance, the misclassification error will increase, which results in a high PI. If the feature only contains redundant information also captured by other features, the misclassification error will not change much and the PI will be low. The proposed feature selection algorithm starts from this standard RF feature selection method, the sections below explain the different improvements that were implemented.

B. Patient independent validation

In many clinical applications feature values might be patient dependent with large inter-patient variabilities. These patient dependencies should be avoided, and the validation performance should take into account the generalizability of the classifier to new, unseen patients. To account for this, the selection of the OOB validation samples in the RF was adapted. Instead of selecting the OOB samples randomly, one third of the patients were randomly selected for each tree, and all samples of these patients were used as OOB samples. As such, the OOB prediction for each sample is only based on the trees that were trained without using any information of that patient, thereby avoiding overfitting.

C. Cohen kappa value as performance estimate

In the standard RF feature selection method, the misclassification error is used to assess the performance. In unbalanced or multi-class problems the Cohen's kappa value is, however, a more appropriate performance measure [9]. The kappa value was designed as a measure of inter-rater agreement and gives an estimate of how good the agreement is between the labels of two raters, or in this case between the gold standard class labels and the predicted classes, with respect to the degree of agreement that can be expected due to chance. This value is computed as

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{1}$$

with p_o the observed agreement and p_e the expected agreement. By comparing with the chance level, the method accounts for unbalanced data and this also enables the comparison of performances between datasets with different prior class distributions. In order to have a minimization problem, as is the case with the misclassification error, the $(1 - \kappa)$ value will be used as performance measure.

D. Inclusion of feature correlation

It is expected that the standard RF feature selection should be able to remove correlated features, since the RF will obtain the information contained in the feature under investigation via the correlated feature. However, studies have shown that the PI tends to be overestimated for correlated features [10]. This might be due to the random sampling of features at each node, which could cause the correlated feature to be excluded. To avoid the selection of redundant features, the predictor importance was multiplied with one minus the maximum absolute Spearman correlation coefficient between the feature under investigation and the other remaining features as follows

$$PI = \Delta\kappa \times (1 - C_{max}). \tag{2}$$

The PI of features that are highly correlated with another feature will then be further decreased, this will lead to more feature pruning. Moreover, to speed up computations, highly correlated features are removed at the beginning of the algorithm. For feature pairs with a Spearman correlation coefficient higher than 0.9 only the feature with the highest F-score was retained.

E. Determination of the optimal number of features

In order to determine what the optimal number of removed features is, the mean OOB Cohen kappa error over all trees were compared. Additionally, this value was multiplied with one minus the average correlation coefficient in the corresponding feature set. The feature set with a maximal value on this kappa-correlation performance parameter was selected (excluding the one with only 1 feature, except if this case would have the highest OOB kappa coefficient as well).

III. Application: sleep apnea detection

The proposed feature selection method was applied to the problem of sleep apnea detection using the SpO_2 signal.

A. Dataset

This study used two datasets for training and testing the method. The first dataset is the publicly available Sleep Heart Health Study (SHHS) [11], [12], which contains PSGs of 5793 general population subjects at baseline (SHHS1) and a follow-up PSG on average 5 years later for 2651 of these subjects (SHHS2). Following [5], 500 recordings of SHHS1 were selected as training set, which were used to perform the feature selection and training of the RF. The remaining recordings were used as independent test sets.

The second dataset was only used for testing and contains data recorded at the sleep laboratory of the University Hospitals Leuven (UZ Leuven). The study was approved by the ethical committee of UZ Leuven and each patient signed an informed consent. This dataset contains PSGs of 218 patients who came to the clinic for a diagnosis of sleep apnea. 91 of them turned out to have an AHI lower than 15.

For both datasets, annotations of start and end points of apneic events were available according to the AASM2012 scoring rules [3] (the SHHS annotations were adapted as described in [5]).

B. Feature extraction

The preprocessing and feature extraction of the SpO_2 signals followed the same workflow as in [5]. All oxygen desaturations were detected in the SpO_2 signal and a total of 143 time-domain features were extracted from each desaturation. All features were log transformed, which resulted in 286 features in total. Desaturations were then linked to annotated apneic events if an event occurred within 60 seconds before the start of the desaturation. A RF classifier was then trained to predict if a desaturation was apnea related or not.

Additionally an estimate of the AHI was computed as the number of apnea related desaturations divided by the total recording time. This estimate was corrected using a robust linear fit on the training set to account for apneas without desaturations and wake time.

IV. Results

A. Feature selection

The feature selection was applied to the obtained training set. The resulting OOB kappa coefficients and kappa-correlation performance parameters are indicated in Fig. 1 according to the number of removed features. The optimal feature set was obtained when 283 features were removed, or in other words when 3 features were retained: the Log amplitude of desaturation, the Log downward phase rectified signal averaging (PRSA) amplitude and the variance of the second order derivative of the SpO_2 signal during desaturation.

B. Classification performance

The performance results of the desaturation classification for the different datasets can be found in Table I. The table compares the results obtained with the improved feature selection to the results reported in [5], where 6 features were included.

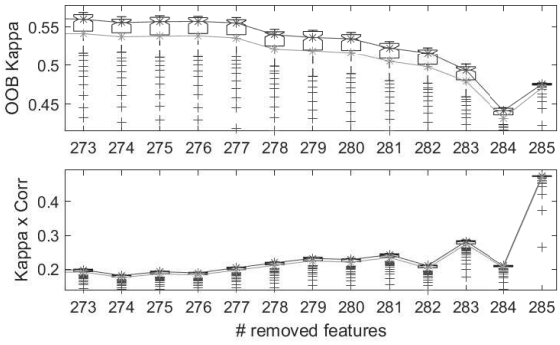

Fig. 1. Overview of classifier performance values, according to the number of features that were removed. On top the OOB Cohen kappa values per tree are plotted. On the bottom these values were multiplied with one minus the mean correlation coefficient in the feature set.

TABLE I

Overview of the desaturation classification performance for the different datasets before and after the new feature selection method.

	Dataset	Acc	Se	Sp	AUC	κ
	SHHS1 train	82.8	64.6	89.3	85.7	0.549
[5]	SHHS1 test	83.3	60.8	90.2	85.0	0.522
	SHHS2	82.0	70.7	85.7	86.1	0.535
6 features	UZ Leuven	79.6	71.4	84.2	84.9	0.557
Current	SHHS1 train	80.8	63.8	86.9	83.2	0.506
study	SHHS1 test	81.6	60.0	88.2	82.2	0.484
	SHHS2	80.6	68.2	84.5	83.7	0.499
3 features	UZ Leuven	76.6	68.9	80.9	81.6	0.496

C. SAHS screening

Table II contains the results for the SAHS screening according to AHI cut-off points of 5, 15 or 30 apneic events per hour of sleep. The results of this study are again compared with those reported in [5].

V. Discussion

The proposed algorithm resulted in the selection of 3 features compared to the 6 features obtained in [5]. Therefore, the number of features was reduced to half, whereas, this only resulted in a minor decrease in classification performance, see Table I. Moreover, the clinical relevance of this drop in performance can be assessed by looking at the SAHS screening parameters in Table II, only slight changes in area under the curve (AUC) were detected, while now only half of the features need to be computed. This greatly increases the interpretability of the classifier, the 3 features can now be visualized in a 3d plot, as is shown in Fig. 2 for one patient. This can help to define why certain desaturations or patients could not be classified correctly.

In [5], the feature selection was performed based on the mRMR algorithm. This algorithm, however, does not provide an optimal number of features to remove. Therefore, an extra backwards wrapper needed to be added in order to remove remaining redundant features. The proposed feature selection

TABLE II

OVERVIEW OF THE APNEA SCREENING PERFORMANCE, WHEN TAKING AN AHI CUT-OFF OF 5, 15 OR 30 FOR THE DIFFERENT DATASETS, BEFORE AND AFTER THE NEW FEATURE SELECTION METHOD.

	Dataset	Acc 5	Se 5	Sp 5	AUC 5	Acc 15	Se 15	Sp 15	AUC 15	Acc 30	Se 30	Sp 30	AUC 30
	SHHS1 train	96.2	97.1	93.6	99.1	96.0	95.6	96.4	99.2	93.0	83.2	96.3	97.1
[5]	SHHS1 test	83.0	81.3	91.3	93.1	87.5	77.4	95.4	95.0	93.1	84.6	94.7	96.9
	SHHS2	89.7	93.1	73.6	94.2	88.1	90.1	86.4	95.3	90.3	92.5	89.9	97.0
6 features	UZ Leuven	81.2	79.8	92.0	93.1	85.3	79.5	93.4	93.6	91.7	89.2	93.1	97.9
Current	SHHS1 train	89.2	90.1	86.4	93.9	91.0	90.4	91.6	96.7	91.8	80.8	95.5	96.4
study	SHHS1 test	84.4	84.2	85.2	92.4	87.4	85.8	88.7	94.5	93.4	85.9	94.9	96.8
	SHHS2	89.8	93.4	72.7	93.8	84.7	92.9	77.8	94.4	91.0	90.5	91.2	96.6
3 features	UZ Leuven	81.2	81.4	80.0	90.3	82.6	80.3	85.7	91.7	89.0	86.5	90.3	96.6

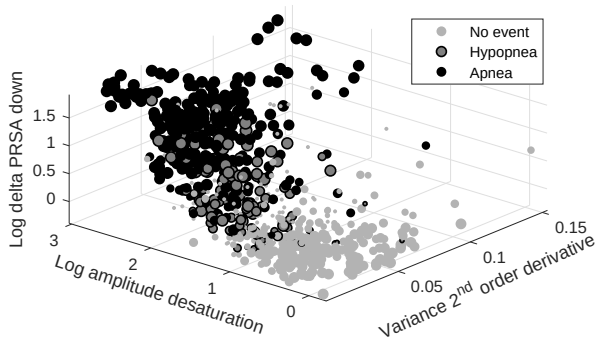

Fig. 2. 3D plot of the selected features for one patient.

method in the current study has the advantage of providing an optimal number of features. Moreover, the algorithm will result in a better pruned feature set, even if it results in a slight decrease in performance.

Since the method is based on the Cohen kappa value, it will be very useful in unbalanced multiclass classification problems. The stability of the Cohen kappa value over datasets with different characteristics can be noticed in Table I. While the accuracy varies between the SHHS and UZLeuven datasets, the kappa values remain stable. Moreover, the patient independent validation by selecting OOB per patient is important in many clinical problems. This adaptation results in a more realistic estimation of the validation performance of the training set. In the sleep apnea example this can be observed in the SAHS screening results in Table II, the training AUCs were around 99%, whereas they are now more in line with with the test performances.

The feature selection method should be evaluated on benchmark datasets with more challenging problems for further evaluation and comparison with existing methods.

For SAHS screening, a new UZLeuven dataset was added compared to [5], which contains patients with various AHI levels. This enables the evaluation of the SAHS screening performance on this clinical dataset containing patients coming to the sleep laboratory for diagnosis of sleep apnea. The results in Table II show that the screening performance is quite stable between the general population train and test sets of the SHHS and the clinical UZLeuven datasets. For an AHI cut-off value of 15 an AUC of 91.7-94.5% can be achieved with an accuracy of 82.6-87.4%, when shifting the cut-off to 30, an AUC of 96.6% is obtained with an accuracy of about 90%.

VI. CONCLUSION

This paper presented a feature selection method based on changes in OOB kappa errors of a random forest classifier, including a feature correlation correction and patient independent validation. This method was applied to the SpO_2 based sleep apnea detection. The number of resulting features was halved with respect to previous feature selection methods, thus improving the interpretability of the classifier, while only a slight decrease in performance was noticed. Moreover the clinical screening performance remained similar.

REFERENCES

[1] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[2] H. Peng et al, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[3] R. B. Berry et al, "Rules for scoring respiratory events in sleep: update of the 2007 aasm manual for the scoring of sleep and associated events," *J Clin Sleep Med*, vol. 8, no. 5, pp. 597–619, 2012.

[4] T. D. Bradley and J. S. Floras, "Obstructive sleep apnoea and its cardiovascular consequences," *The Lancet*, vol. 373, no. 9657, pp. 82–93, 2009.

[5] M. Deviaene et al, "Automatic screening of sleep apnea patients based on the spo_2 signal," *IEEE Journal of Biomedical and Health Informatics*, 2018.

[6] J. Lázaro et al, "Pulse rate variability analysis for discrimination of sleep-apnea-related decreases in the amplitude fluctuations of pulse photoplethysmographic signal in children," *IEEE journal of biomedical and health informatics*, vol. 18, no. 1, pp. 240–246, 2014.

[7] M. Deviaene et al, "Sleep apnea detection using pulse photoplethysmography," in *Proceedings of the 45th Annual Computing in Cardiology Conference*. Wordpress, 2018, pp. 1–4.

[8] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[9] A. Ben-David, "Comparison of classification accuracy using cohens weighted kappa," *Expert Systems with Applications*, vol. 34, no. 2, pp. 825–832, 2008.

[10] C. Strobl et al, "Conditional variable importance for random forests," *BMC bioinformatics*, vol. 9, no. 1, p. 307, 2008.

[11] S. F. Quan et al, "The sleep heart health study: design, rationale, and methods," *Sleep*, vol. 20, no. 12, pp. 1077–1085, 1997.

[12] S. Redline et al, "Methods for obtaining and analyzing unattended polysomnography data for a multicenter study," *Sleep*, vol. 21, no. 7, pp. 759–767, 1998.