CrossMark

Random forest-based approach for physiological functional variable selection for driver's stress level classification

Neska El Haouij[1,2,3] · Jean-Michel Poggi[2,4] ·
Raja Ghozi[1] · Sylvie Sevestre-Ghalila[3] ·
Mériem Jaïdane[1]

Abstract This paper deals with physiological functional variables selection for driver's stress level classification using random forests. Our analysis is performed on experimental data extracted from the *drivedb* open database available on PhysioNet website. The physiological measurements of interest are: electrodermal activity captured on the driver's left hand and foot, electromyogram, respiration, and heart rate, collected from ten driving experiments carried out in three types of routes (rest area, city, and highway). The contributions of this work touch on the method as well as the application aspects. From a methodological viewpoint, the physiological signals are considered as functional variables, decomposed on a wavelet basis and then analyzed in search of most relevant variables. On the application side, the proposed approach provides a "blind" procedure for driver's stress level classification, giving

✉ Neska El Haouij
  elhaouij.nsk@gmail.com

  Jean-Michel Poggi
  Jean-Michel.Poggi@math.u-psud.fr

  Raja Ghozi
  raja.ghozi@enit.utm.tn

  Sylvie Sevestre-Ghalila
  Sylvie.Sevestre-Ghalila@cea.fr
  http://www.cea-tech.fr

  Mériem Jaïdane
  meriem.jaidane@enit.utm.tn

1 National Engineering School of Tunis, University of Tunis El Manar, Tunis, Tunisia

2 Paris-Sud University, Orsay, France

3 CEA-Tech (CEA-LinkLab), CEA, Tunis, Tunisia

4 Paris Descartes University, Paris, France

close performances to those resulting from the expert-based approach, when applied to the *drivedb* database. It also suggests new physiological features based on the wavelet levels corresponding to the functional variables wavelet decomposition. Finally, the proposed approach provides a ranking of physiological variables according to their importance in stress level classification. For the case under study, results suggest that the electromyogram and the heart rate signals are less relevant compared to the electrodermal and the respiration signals. Furthermore, the electrodermal activity measured on the driver's foot was found more relevant than the one captured on the hand. Finally, the proposed approach also provided an order of relevance of the wavelet features.

Keywords Physiological signals · Functional data · Random forests · Recursive feature elimination · Wavelets · Grouped variable importance

Mathematics Subject Classification 62H30 · 62P30

1 Introduction

The goal of this study is to classify stress levels of drivers in real-world driving situations. To this end, a random forest-based method is used for the selection of physiological functional variables captured during driving sessions. First, we present the "affective computing" context of our work. Then, we give an overview of previous studies that used the *drivedb* database of physiological measurements. Subsequently, we describe our method for functional data analysis and variable selection. Finally, we discuss the importance of grouped variables.

Stress recognition during driving

In recent years, the field of affective computing emerged as an interdisciplinary field combining computer science and engineering with cognitive science, physiology, and psychology. This field is expected to be beneficial for improving products in the automotive industry in order to enhance the quality of the driving experience and the driver's comfort (Bostrom 2005). Actually, according to the American Highway Traffic Safety Administration, high stress levels negatively impact drivers reactions, especially in critical situations (Smart et al. 2005), which is one of the lead causes in vehicle accidents. Hence, the ability to detect driver's stress in a timely fashion may provide key indicators for valuable and timely decision making. In addition, a driver's stress level monitoring is important in order to avoid traffic accidents and hence promote safe and comfortable driving. For instance, researches have used sensors for data collection in order to build models for better human affective state understanding and to design smart human-centric interfaces (Tao and Tan 2005).

Physiological data such as Electrocardiogram (ECG), Electromyogram (EMG), Electrodermal Activity (EDA) and Respiration (Resp) have been proven to reflect stress levels during driving tasks (Singh and Queyam 2013; Singh et al. 2012; Deng et al. 2012; Rigas et al. 2008). Usually, various types of features are extracted from these physiological measurements. It is thus crucial to select only features that are

relevant in the recognition of the different stress levels during a given task (such as driving in different routes). In this context, there is a dearth of literature that focused on features selection and the study of the correlation between the selected features and the driving route complexity (Singh and Queyam 2013).

In practice, given that real-world driving data collection is very challenging, having access to an open database with physiological measurements captured while driving, offers great benefits in carrying studies such as the one at hand. Unfortunately, very few open databases are available. We have used in our study the open database "*drivedb*", available on the PhysioNet website (Goldberger et al. 2000). The data were captured during real-world driving experiments that were originally collected by Healey (2000). The database includes various physiological measurements namely: ECG, EMG, EDA (measured on hand and foot), and respiration of 17 drives performed along a fixed itinerary, which switches between city and highway driving conditions. The original study analyzed twenty two features, derived from non-overlapping segments of physiological signals, extracted from the rest, highway, and city periods of all drives (Healey and Picard 2005). A correlation analysis was performed between the features extracted from the physiological signals and the driver's affective state assessed by a stress metric created from the experiment video recording. We recall that the original study of Healey and Picard (2005) found that the driver's stress level was correlated the most with the EDA and the Heart Rate (HR) extracted from the ECG. However, that study did not formally address the selection of physiological features. Moreover, it did not consider any use of multi-resolution analysis of any physiological signal.

It is worth noting that since then, several studies have relied on the *drivedb* database. We note in particular (Sidek and Khalil 2011; Chaudhary 2013; Karmakar et al. 2014) studies that used ECG signals while (Imam et al. 2014) used ECG and Respiration data. Several studies on stress level recognition relied on the complete set of the physiological variables: we note in particular the work of Zhang et al. (2010) that proposed an approach based on Bayesian Network for the fusion of features extracted from the whole *drivedb* database. Akbas (2011) presented an evaluation of the mean and standard deviation computed on the various *drivedb* physiological signals, covering all rest and driving periods of the experiments. In Deng et al. (2012) study, the importance of features selection on stress level recognition was explored using the whole data of *drivedb*. We note that the bibliographical study of Sharma and Gedeon (2012) found that the stress modeling techniques did not include Random Forests (RF). Recently, Guendil et al. (2015); Ayata et al. (2016); Granero et al. (2016) used RF as a classification technique for emotion recognition. But there is no study that benefited from the variable importance offered by RF in order to select or rank the physiological variables allowing stress level recognition.

Variable selection and functional data

We recall that the random forests algorithm, introduced by Breiman (2001), is based on aggregating a large collection of tree-based estimators. These methods have relatively good predictive performances in practice and work well in high dimensional problems. The RF power was demonstrated in several studies, summarized in Verikas et al.

(2011). Moreover, random forest methods provide several measures of the importance of the variables with respect to the prediction of the outcome variable. It has been shown that the permutation importance measure introduced by Breiman, is an efficient tool for variables selection (Díaz-Uriarte and Andrés 2006; Genuer et al. 2010; Gregorutti et al. 2016). One of the main issues of variable selection methods in high dimensional data (data with small number of observations compared to the number of variables) is their instability. Indeed, the set of selected variables may change when slight perturbation is introduced to the training set. One solution to solve this instability consists in using bootstrap samples and a stable solution can then be obtained by aggregating selections performed on several bootstrap subsets of the training data. A similar solution is in fact proposed in this paper by introducing a selection through repetitions of random forests based methods, intensively based on bootstrap resampling.

Functional Data Analysis (FDA) is a field that analyzes data for which explanatory variables are functions (of time in our case). One possible approach in FDA consists in considering the projections of the observations into finite dimensional spaces such as splines, wavelets, Fourier (see for instance Ramsay and Silverman 2005; Ferraty and Vieu 2006). Many studies (Ullah and Finch 2013; Ramsay and Silverman 2002) propose classification or regression methods for functional data in one of two possible situations: with one functional predictor and recently for several functional variables. Classification based on several possibly functional variables has also been considered using the CART algorithm for similar driving experiences in the study of Poggi and Tuleau (2007), and using SVM in Yang et al. (2005). Variable selection using random forests was recently performed in the study of Genuer et al. (2015). In our study, we will adopt multivariate functional data using random forests and the grouped variable importance measure proposed by Gregorutti et al. (2015).

The main contributions of this study are twofold: on the methodological side, it takes advantage of the functional nature of the physiological data and offers a procedure of data processing and variable selection. For that, the physiological signals are decomposed on a common wavelet basis and then analyzed in search of important variables using grouped variable importance. This analysis is applied on two levels of data selection strategies, based on a proposed "endurance" score. On the applied side, the proposed method provides a blind (i.e. without prior information) procedure of driver's stress level classification that does not depend on the extraction of expert-based features of physiological signals.

Paper outline

This paper is organized as follows. After this introductory section, Sect. 2 is dedicated to the description of the protocol and the database used in this study. Section 3 recalls the random forest model and the variable selection procedure based on the variable importance measure. Section 4 presents the three main steps of variables selection. Results of the variables selection applied to the "drivedb" database are presented in Sect. 5. Finally, Sects. 6 and 7 present the discussion and the conclusions of the work.

2 Experimental protocol and data collection

The *drivedb* database was selected in our study for two main reasons: it is an open access database and its content has been explored in several studies due to its clear annotation. In fact, each dataset related to a drive includes a marker signal providing annotation of each driving segment, namely: city, highway driving, and the rest periods. In this section, we present the real-world driving protocol, originally proposed by Healey (2000). Specifically, we will describe the cohort and detail the data construction process.

2.1 Real-world driving protocol

All the driving experiments were based on a drive path which extended over 20 miles of open roads in the greater Boston area (Healey and Picard 2005). The driver was asked to follow a set of instructions in order to keep the drive regular. He/she was requested to respect the speed limits and to shut down the radio. The choice of the driving path was considered to imitate a typical daily commute and thus would induce reactions revealing stress levels usually evoked in normal daily driving trip. To ensure that all drivers made the same route, they were shown a map of the driving path before the driving session. An observer accompanied each participant on his/her drive. The observer sat in the rear seat in order not to interfere with the driver's normal behavior and affective state. This observer has also to monitor the experiment using a laptop displaying the recorded physiological signals and videos recording the inside and the outside scenes of the vehicle. Depending on the traffic density, the drive duration varies between 50 and 90 min. All driving sessions began with a 15 min rest period for the driver, after which the driver drove the car out of the garage, onto a ramp until he/she reached a congested city street. The driver then went into a main street, which has several traffic signals. It is assumed that driving on this road produces high stress-levels. The path continued from the city into a highway, where medium stress levels were usually evoked. After exiting the highway, the driver turned around, re-entered the highway in the opposite direction, followed the same route, and got back to the garage. After a complete driving session, the driver was requested to rest again. For each of the rest periods, the driver is asked to sit in the garage inside the vehicle (with an idle engine) with closed eyes. Thereby, this setting inducing the lowest stress level along the experiment was introduced to establish baseline measurements. Each driving session includes periods of rest, highway, and city driving (see Fig. 1), meant to produce respectively low, medium, and high stress levels. These assumptions were validated by two methods developed in Healey and Picard (2005): the first method included a survey and the second consisted on a score derived from observable events and actions coded from the video tapes recorded during each drive.

2.2 Cohort description

The set of data used in our work refers to 10 drives carried out by 4 participants denoted by M-3, M-4, F-8 and Ind 4, whose individual driving sequence details are shown in
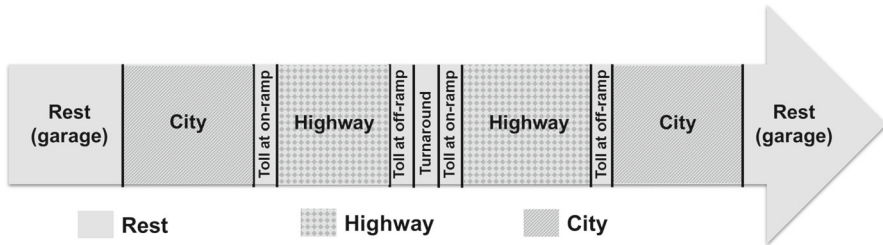
Fig. 1 Description of the driving setting. For simplification, the toll and the turnaround segments are considered as city driving

	Drive	Participant label	Date (mm-dd-yy)	Duration (hh:mm:ss)
Table 1 Description of the different 10 drives. Note that each participant is labeled by a sequence composed of gender (M or F) followed by the number of years of the driving experience. No information was available on Ind 4	1	M-3	07-28-99	1:24:15
	2		08-04-99	1:20:46
	3	M-4	07-15-99	1:28:38
	4		08-05-99	1:21:11
	5		08-13-99	1:10:52
	6	F-8	08-02-99	1:21:16
	7		08-05-99	1:21:13
	8		08-06-99	1:23:04
	9		08-09-99	1:17:38
	10	Ind 4	07-16-99	1:04:57

Table 1. According to the original study of Healey (2000), participant M-3, contributed in the database by 2 drives and was an undergraduate male. He had 3 years of driving experience though he was not a regular driver. Participant M-4 is an undergraduate student with over 4 years of driving experience. He had not driven a month prior to the experiment. He contributed to the database by 3 driving experiments. Participant F-8 was a female undergraduate having 8 years of driving experience. Four driving experiments in the database are related to this driver. Finally, Ind 4 participated by just one driving experience. No gender or driving experience information of Ind 4 were provided.

2.3 Data construction

The open database *drivedb*, available in PhysioNet website, contains only 17 out of the 24 total driving experiments (Drive01, Drive02, …, Drive16, Drive17a, Drive17b).

In our study, only 10 out of those 17 driving experiences are considered (namely: Drive05, Drive06, …, Drive12, Drive15, Drive16), since not all data were complete for the listed drives. In fact, seven out of the 17 drives report incomplete data including missing information, as indicated in Table 2.

For each of the 10 considered drives, seven segments, of 5 min each, were extracted for analysis in our work, as shown in Fig. 2. However, two 5-min segments (correspond

Table 2 Details of the missing data for the 7 drives excluded from the analysis due to the reported incomplete data in the *drivedb* database

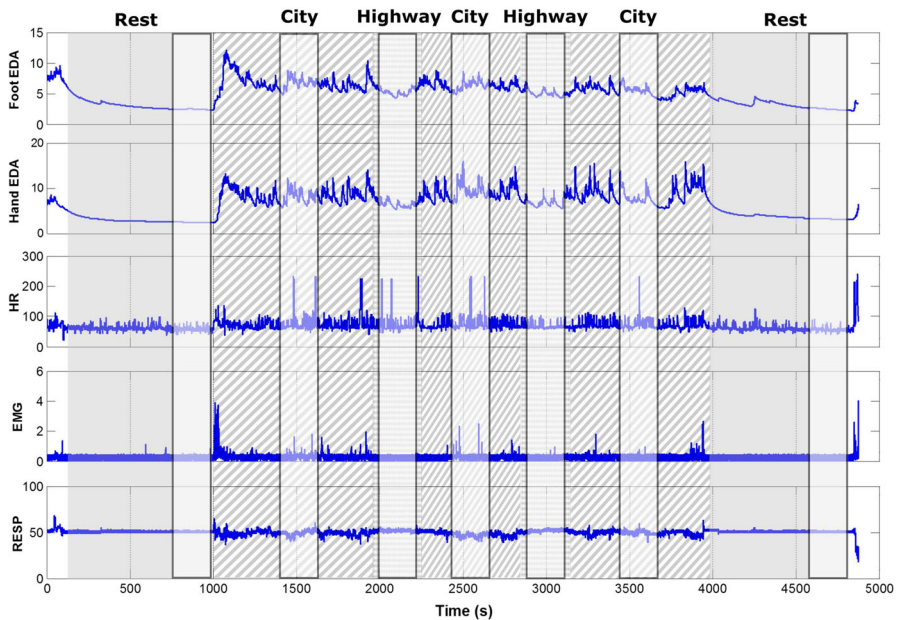| Drive number	Missing data
01	Marker
02	Hand EDA and EMG
03	HR, EMG and Marker
04	EMG
13	Hand EDA
14	HR
17a and 17b (two parts of one experiment)	Marker is not complete

Fig. 2 Illustration of segment extraction of different physiological signals of Drive 7. Note that the physiological data were stored based on the same sampling frequency $F_s = 15.5$ Hz

to the second rest period of Drive 5 and Drive 10) were not considered since the markers that annotate the end of the second rest period are missing. Thus only 68 segments were subject of our analysis. In order to avoid driver's both memory recall and anticipatory effects, only the middle segments of highway and city driving are retained as representative samples of respectively medium and high stress levels. This choice was proposed by Healey and Picard (2005) where also the last 5 min of the two rest periods are considered, and this was chosen in order to give the participant enough time to relax from the previous tasks, thus inducing a low stress baseline.

Since *drivedb* data is in raw form, a standardization task is performed. For that, we adopt the same preprocessing proposed by Healey and Picard (2005), were the data were normalized in order to ensure the comparability between individuals.

For instance, EDA standardization, proposed by Lykken (1972), is achieved by centering the EDA values using the minimum value (during the first rest period), then dividing by the EDA overall range (i.e. the difference between the maximum of the signal during the entire experiment and the minimum value during the first rest period). The HR, EMG and RESP signals are normalized by subtracting from each sample the mean value of the each signal over the first rest period. The physiological data were stored based on one same sampling frequency $F_s = 15.5$ Hz.

Let S be the vector of physiological signals used here as explanatory variables. We extract 5 min segments. The choice of this window size allows the extracted segments to be informative and comparable. Five minutes corresponds to 4650 samples based on the used sampling period $\Delta t = \frac{1}{F_s} = \frac{1}{15.5} = 0.065$ s.

In order to project these segments into a wavelet basis (see Sect. 4.1), we keep the first 2^{12} samples (since a power of two simplifies the discrete wavelet decomposition) which corresponds to 4096 samples. Functions $S = (S^1, \ldots, S^p) \in \mathcal{S}^p$ where $S^j = \{S^j(t) \in \mathbb{R}, t \in [0, T]\} \in \mathcal{S}$ can be considered as functions of finite energy over $[0, T]$ where $T = 264.26$ s and $j = 1, \ldots, p$ where $p = 5$.

Let i be the index of the extracted segment, $i = 1, \ldots, N$ where $N = 68$.

For a given i, $S_i(t) = (S_i^1(t), \ldots, S_i^p(t))$ presents 5 physiological signals used as explanatory variables corresponding to the stress level y_i,

$$y_i = \left\{ \begin{array}{l} \text{H} = \text{High stress level,} \\ \text{M} = \text{Medium stress level,} \\ \text{L} = \text{Low stress level.} \end{array} \right\}$$

3 Random forests and variables selection using variable importance

3.1 Random forests and variable importance measure

Let $X = (X^1, \ldots, X^p)$, $X \in \mathbb{R}^p$ denote the explanatory variables and Y the response variable that takes numerical values in the case of regression context, and a class label in the case of classification.

Let $L = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be a learning set, consisting in n independent observations of the vector (X, Y).

Random Forests (RF) is a non-parametric statistical method, originally introduced by Breiman (2001) as an extension and improvement of decision trees (Breiman et al. 1984). RF provides estimators of either a Bayes classifier ($f : \mathbb{R} \mapsto \mathcal{Y}$ minimizing the classification error $P(Y \neq f(X))$), in the classification case, and of the regression function g that verifies $Y = g(X) + \varepsilon$ with $E[\varepsilon|X] = 0$, in the regression case (Hastie et al. 2001). A random forest is basically a set of decision trees, constructed over n_{tree} bootstrap samples $L^1, \ldots, L^{n_{tree}}$ of the training set L.

At each node, a fixed number of variables is randomly picked to determine the splitting rule among them. The trees are not pruned so all the trees of the forest are maximal trees. The resulting learning rule is the aggregation of all the estimators resulting from those trees, denoted by $\hat{f}_1, \ldots, \hat{f}_{n_{tree}}$. The response value at a new point x is hence obtained by the aggregation which consists in building the following

$$\hat{f}(x) = \begin{cases} \frac{1}{n_{tree}} \sum_{k=1}^{n_{tree}} \hat{f}_k(x) & \text{in regression,} \\ arg\,max_{1 \leq c \leq C} \sum_{k=1}^{n_{tree}} 1_{\hat{f}_k(x)=c} & \text{in classification.} \end{cases} \quad (1)$$

We now introduce the Out-Of-Bag (OOB) sample that will be used to define the variable importance measure, provided by the RF. For each tree t, the OOB sample, denoted by OOB_t, is the set of observations that are not included in the bootstrap sample used to construct t. This OOB sample leads to a "smart" estimate of the error by calculating the error of tree t on the observations of OOB_t which can be used as test sample. This estimate is called the OOB error of a tree t. In order to assess the contribution of a variable to explain the response variable of interest in the RF model, we can use the permutation importance measure originally proposed by Breiman (2001), and denoted henceforth as VI. It is defined as the mean increase, over all the trees in the forest, of the OOB error of a tree obtained when randomly permuting the variables in the OOB samples. The OOB error of a tree is measured by the Mean Square Error (MSE) for the regression case and by the misclassification rate for the classification on the OOB sample.

More formally, a variable X^j is considered important if when breaking the link between X^j and Y, the prediction error increases. The prediction error of each tree $\hat{f}$ is evaluated among its OOB sample with the empirical estimator

$$\hat{R}(\hat{f}, \bar{L}) = \begin{cases} \frac{1}{|\bar{L}|} \sum_{i:(X_i,Y_i)\in\bar{L}} (Y_i - \hat{f}(X_i))^2 & \text{in regression,} \\ \frac{1}{|\bar{L}|} \sum_{i:(X_i,Y_i)\in\bar{L}} 1_{\hat{f}(X_i)\neq Y_i} & \text{in classification.} \end{cases} \quad (2)$$

Let $\left\{ \bar{L}_k^j, k = 1, \ldots, n_{tree} \right\}$ refers to the set of OOB permuted samples resulting from random permutations of the values of the j-th variable in each out-of-bag sample OOB_{t_k}.

The VI is defined as the mean increase in the prediction error (estimated thanks to the OOB error) over all trees, and estimated by

$$\hat{I}(X^j) = \frac{1}{n_{tree}} \sum_{k=1}^{n_{tree}} \left[\hat{R}\left(\hat{f}_k, \bar{L}_k^j \right) - \hat{R}\left(\hat{f}_k, \bar{L}_k \right) \right]. \quad (3)$$

Several authors were interested in the numerical study of this VI (see Strobl and Zeileis 2008; Nicodemus et al. 2010; Auret and Aldrich 2011). Some theoretical results are also available in this regard (see Louppe et al. 2013; Gregorutti et al. 2015). For instance, Zhu et al. (2012) proved that $\hat{I}(X^j)$ converges to $I(X^j)$, where

$$I(X^j) = E[(Y - f(X^{(j)}))^2] - E[(Y - f(X))^2]. \quad (4)$$

where $X^{(j)} = (X^1, \ldots, X'^j, \ldots, X^p)$ is a random vector such that X'^j is an independent replication of X^j, independent of Y and of X^k, $k \neq j$.

The selection of groups of variables is a relevant topic in itself, see for example the Group Lasso allowing to select groups of variables in the linear model using the lasso selection strategy (see Bach 2008). In our case, obvious interesting groups of

variables are, for a given functional variable, the coefficients of the function considered as a whole or the wavelet coefficients at a given level, for example. In order to generalize the random forests based selection strategies, Gregorutti et al. (2015) extend the permutation importance definition for a group of variables. To estimate the permutation importance of a group of variables denoted by $\mathbf{X^J}$, let us consider for each $k \in \{1, \ldots, n_{tree}\}$, $\bar{\mathbf{L}}_k^J$ the permuted set of $\bar{\mathbf{L}}_k$ resulting by randomly permuting the group $\mathbf{X^J}$ in each OOB sample $\bar{\mathbf{L}}_k^j$. The permutation importance of $\mathbf{X^J}$ is estimated by

$$\hat{I}(\mathbf{X^J}) = \frac{1}{n_{tree}} \sum_{k=1}^{n_{tree}} \left[\hat{R}\left(\hat{f}_k, \bar{\mathbf{L}}_k^J\right) - \hat{R}\left(\hat{f}_k, \bar{\mathbf{L}}_k\right) \right]. \tag{5}$$

When the number of variables considered in a group increases, the grouped VI may increase (Gregorutti et al. 2015), which is particularly true in the case of an additive model and independent variables. In order to take group size into account, we shall use the normalized version of the VI (Gregorutti et al. 2015):

$$I_{nor}(\mathbf{X^J}) = \frac{1}{|J|} I(\mathbf{X^J}). \tag{6}$$

3.2 Variable selection using random forest-based recursive feature elimination

Random Forests-based Recursive Feature Elimination (RF-RFE) algorithm, summarized in Table 3, was originally proposed by Gregorutti et al. (2015), and inspired from Guyon et al. (2002) that introduced the Recursive Feature Elimination algorithm for SVM (SVM-RFE).

Let us first sketch the statistical model used in the Random Forest framework. Let $L = \{(S_1, Y_1), \ldots, (S_n, Y_n)\}$ be a learning set, consisting in n independent observations of the random vector (S, Y). The $(S_i(t), y_i)$, introduced in Sect. 2.3, are then considered as realization of unknown distribution of this sample. We aim to build an estimator of the Bayes classifier $f : \mathbb{R} \mapsto \mathcal{Y}$ that minimizes the classification error

Table 3 Summary of the selection algorithm based on RF-RFE

1. Split the whole data L into a training set L_T containing $\frac{2}{3}$ of the data and a validation set L_V containing the remaining $\frac{1}{3}$. Set the subset of the variables $\mathcal{V}$ to the whole explanatory variables

2. Fit a random forest model using L_T and considering the set of variables $\mathcal{V}$

3. Compute the VI measure (respectively the grouped VI measure)

4. Compute the error using the validation sample L_V

5. Eliminate the least important variable (resp. group of variables) and update $\mathcal{V}$

6. Repeat 2–5 until no further variables (resp. groups of variables) remain

7. Select the variables (resp. the groups of variables) involved in the model minimizing the prediction error

$P(Y \neq f(S))$. We denote by $\hat{f}$ an estimator belonging to the fully non parametric set of models given by the random forests framework introduced in Sect. 3.1.

The proposed algorithm in Table 3 will be briefly explained in the case of scalar variables since the case of functional variables is similar. In fact, the variable groups are predefined depending on the type of data, and in our case this definition depends on the functional reconstruction procedure. At the first step, the dataset is randomly split into a training set containing two thirds of the data and a validation set containing the remaining one third. Steps 2–5 of the procedure are a loop. This loop starts by fitting the model using all the explanatory variables using Random Forests. Then, the variables are ranked from the most to the least important based on their VI values, computed on the training set. Then, the least important predictor is eliminated, the model is re-trained and the prediction error computed on the validation set. The variable ranking and elimination steps are repeated until no variable remains. The final model is chosen by minimizing the prediction error. It should be noted that at each iteration, the predictors importance is recomputed according to the model composed of the reduced set of explanatory variables. Indeed, variations to this approach are conceivable, namely using an ascending strategy where the variables are sequentially invoked according to a given ranking of their respective importance obtained once and for all at the beginning. These variants are suboptimal but less time consuming, which can be especially interesting in the high dimensional case. This exact same algorithm detailed in Table 3 can be applied for the selection of functionnal variables, as well as wavelet coefficients grouped by wavelet level.

4 The 3 steps of the variable selection approach

One goal of the affective computing research in the automotive field is to recognize the driver's stress level. Although using many sensors to capture physiological signals is likely to increase accuracy of the results, but integrating many sensors in the vehicle may be a complex task. In fact, there is a trade-off between the adequate number of sensors and stress level recognition accuracy.

Consequently, we propose an approach that not only removes irrelevant physiological variables from the classification model, but also selects the most relevant wavelet features among the wavelet levels resulting from the decomposition of the retained variables; allowing an optimal stress-level recognition.

A summary of the 3-step procedure is listed in Table 4. This procedure repeatedly applies the RF-RFE algorithm described in Table 3. A detailed description of our approach is given below.

4.1 Step 1. Wavelet decomposition of the physiological functional variables

For the sake of clarity, we shall first recall few basic concepts related to wavelets and their relationship to functional variables. Given a space of functions $\mathcal{F}$ (typically, $L^2([0, 1])$ which is the space of square integrable functions defined on $[0, 1]$) and a probabilistic space Ω, a functional random variable is a measurable application

Table 4 Summary of the proposed 3-step approach

Step 1.	Wavelet decomposition of the physiological functional variables
Step 2.	Physiological functional variable removal:

1. Repeat 10 times: RF-RFE $(G(1), \ldots, G(p))$

2. Compute an "endurance" score for each group $G(\ell)$ as following:

$$score(G(\ell)) = \sum_{m=1}^{p} nbocc\,(G(\ell), Q_m) \times [(p - m) + 1]. \quad (7)$$

where $1 \leq \ell \leq p$, Q_m is the list of the selected variables, all over the 10 executions, at the rank m, $m = 1, \ldots, p$, and $nbocc(G(\ell), Q_m)$ corresponds to the number of occurrences of $G(\ell)$ in the list Q_m

3. Remove the less "enduring" functional variables (those of a score below a threshold δ)

Step 3.	Selection of wavelet levels among the R retained functional variables:

1. Repeat 10 times:

RF-RFE $(\{G(1, S_1), \ldots, G(J, S_1), \ldots, G(1, S_R), \ldots, G(J, S_R)\})$

2. Compute a selection score for each group $G(j, S_r)$ as following:

$$score(G(j, S_r)) = \sum_{m=1}^{p} nbocc\,(G(j, S_r), Q_m) \times [((J \times R) - m) + 1]. \quad (8)$$

where $1 \leq j \leq J$, Q_m is the list of the selected wavelet levels, all over the 10 executions, at the rank m, $m = 1, \ldots, p$, $p = J \times R$, and $nbocc\,(G(j, S_r), Q_m)$ corresponds to the number of occurrences of $G(j, S_r)$ in the list Q_m

3. Select the most "enduring" wavelet levels (those having a score higher than a threshold δ')

$S : \Omega \rightarrow \mathcal{F}$. A function of $\mathcal{F}$ is defined on $[0, 1]$ ($[0, T]$ equivalently in our specific situation) with values in $\mathbb{R}$.

Starting from compactly supported scaling function ϕ and mother wavelet ψ, the sequence of functions $\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k)$ and $\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)$ are obtained by dyadic translations and dilations of ϕ and ψ. This allows us to build several orthonormal wavelet bases, for any $j_0 \geq 0$

$$\mathcal{B} = \left\{\phi_{j_0,k}, k = 0, \ldots, 2^{j_0} - 1\right\} \bigcup \left\{\psi_{j,k}, j \geq j_0, k = 0, \ldots, 2^j - 1\right\}. \quad (9)$$

Therefore, a wavelet transform of a function s of $L^2([0, 1])$ can be written as an expansion on $\mathcal{B}$:

$$s(t) = \sum_{k=0}^{2^{j_0}-1} c_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{\infty} \sum_{k=0}^{2^{j_0}-1} d_{j,k}\psi_{j,k}(t). \quad (10)$$

with $c_{j,k} = \langle s, \phi_{j,k}\rangle_{L^2}$ and $d_{j,k} = \langle s, \psi_{j,k}\rangle_{L^2}$.

The first term in Eq. (10) is the smooth approximation of s at level j_0 while the second term is the detail part of the wavelet representation, then the $c_{j,k}$ are the scaling coefficients and $d_{j,k}$ are the wavelet coefficients at level j and position k. Thanks to the formalism of multi-resolution analysis of L^2, introduced by Mallat (1989), the whole

space is the limit of a sequence of nested subspaces V_j of approximation associated to the scale levels $j \in \mathbb{Z}$ and of increasing resolution. It allows thus to define the Discrete Wavelet Transform (DWT) as a simple algorithm to perform for an input signal of length $N = 2^J$ and for the maximum level of wavelet given by the size N of the sampling grid. The wavelet decomposition of s given in a similar form as Eq. (10) is thus as follows,

$$\tilde{s}_J(t) = c_0\phi_{0,0}(t) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(t). \tag{11}$$

where c_0 is the single scaling coefficient and $d_{j,k}$ denotes the empirical wavelet coefficients derived from applying the DWT to the sampled values.

For a given wavelet function, the physiological functional variables are decomposed into a common wavelet basis at the first step of the approach.

4.2 Step 2. Physiological functional variable removal

The wavelet coefficients provided by the first step of the proposed approach are considered as inputs, which are then grouped by physiological variable. Let $G(\ell)$ denote the group of the wavelet coefficients resulting from the wavelet decomposition of the physiological variable S_ℓ. The RF-RFE algorithm (see Table 3), is applied then on these groups of wavelet coefficients $G(\ell)$ where $1 \leq \ell \leq p$, p is the number of the physiological variables considered in the initial model. As a result, the list of the physiological variables ordered using the VI measure and the number of the selected variables are provided.

It should be mentioned that the values of the results vary for different execution repetitions. Therefore, a procedure is proposed to improve the stability of the method. This procedure consists in executing the RF-RFE 10 times and computing a selection score that helps, at this step, to remove the less "enduring" physiological variables. The "endurance" of a variable measures the ability for it to be selected in the first ranks. The "endurance" score is built based on the number of occurrences of the variable in the lists of selected variables, and on the weights attributed according to the selection rank (see Table 5 for more details on the weights). This score aggregates information concerning the results of 10 executions of the RF-RFE algorithm over the entire set of variables.

Table 5 illustrates this result. For each execution e, the algorithm offers a list of the ranked variables $G(.)_e^1, \ldots, G(.)_e^p$ and a number of the selected variables $nbsel_e$. Let us define Q_m as a list that aggregates the selected variables, all over the 10 executions, at the rank m where $1 \leq m \leq p$. For instance, if $nbsel_e \geq 1$ for an execution e ($1 \leq e \leq 10$) then Q_1 contains the list of the 10 selected variables at the first rank. Q_m may be empty in the case that no variable was selected at the rank m and it may contain several occurrence of the same variable. Let $nbocc(G(.), Q_m)$ denote the number of occurrences of a variable $G(.)$ in the list Q_m.

The "endurance" score of a variable is the weighted sum of the number of occurrences of the variable all over the lists Q_m, $1 \leq m \leq p$. The weight of a variable in a

Table 5 Summary of results of the 10 executions of the RF-RFE

Execution Num	Weights				Number selected var
	p	$p-1$	$\ldots$	1	
Execution 1	$G(\cdot)_1^1$	$G(\cdot)_1^2$	$\ldots$	$G(\cdot)_1^p$	$nbsel_1$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
Execution 10	$G(\cdot)_{10}^1$	$G(\cdot)_{10}^2$	$\ldots$	$G(\cdot)_{10}^p$	$nbsel_{10}$

list of rank m is equal to $(p - m) + 1$. The computation of this score on the physiological variables is detailed in Step 2 of Table 4. The obtained selection scores allow to rank physiological variables according to their "endurance" in the selection procedure. Those having a score less than a threshold δ are removed. Thus, R physiological variables are retained and to be considered in the next step of the procedure.

4.3 Step 3. Wavelet levels selection

This step consists in applying the same procedure detailed in Sect. 4.2, except that for this step the wavelet coefficients are grouped by wavelet levels. In addition, the coefficients considered here concern only the R retained physiological functional variables after the removal process is applied. Let S_r, $1 \leq r \leq R$, be the r-th physiological variable retained after removal based on Step 1. Let $G(j, S_r)$ denote the group of wavelet coefficients resulting from the decomposition of the variable S_r on the wavelet scale level j, $1 \leq j \leq J$ and J is the common maximum scale level that physiological variables S_r, $r = 1, \ldots, R$ can be decomposed on.

Let $nbocc(G(j, S_r), Q_m)$ denote the number of occurrences of the group of wavelet coefficients $G(j, S_r)$ in the list Q_m, where $1 \leq m \leq p$. In this case, $p = R \times J$. The "endurance" score in this step is computed in the same way as in Step 1. It is the weighted sum of the number of occurrences of the group of wavelet coefficients $G(j, S_r)$ all over the lists Q_m, $1 \leq m \leq p$. This score provides then a ranking of the wavelet levels according to their "endurance" in the selection procedure. This step aims mainly to select the most "enduring" wavelet levels. Thus, wavelet levels having an endurance score higher than δ' are selected.

Let us denote the procedure detailed at the Sect. 3.2 as RF-RFE $(Grp_1, \ldots, Grp_p)$ then:

- the procedure described in Sect. 4.2 is indicated with $Grp_\ell = G(\ell)$, corresponding to wavelet coefficients grouped by physiological variables ℓ, $\ell = 1, \ldots, p$.
- the procedure described in Sect. 4.3 is indicated with $Grp_\ell = G(j, S_r)$, corresponding to wavelet coefficients grouped by wavelet scale j, $j = 1, \ldots, J$ of the selected functional variable S_r, $r = 1, \ldots, R$.

Remark It must be noted that the removal of the functional variables (respectively wavelet levels) depends on the choice of the threshold δ (respectively δ'). A classical strategy is to apply a rule similar to the elbow criterion that allows to determine the

number of factors in the PCA, see Jolliffe (2012). We propose here a classical trick: we look at the consecutive differences of the selection score. When starting from the highest values of the endurance scores, we select variables just before the biggest gap or when starting from the lowest values of endurance scores, remove variables just after the big gap. The threshold can then be selected based on a visual inspection, if the number of variables is small.

5 Variables selection on the *drivedb* database

In this section, we present the different variable selection strategies and their applications on the *drivedb* database. The objective of variable selection is not only to remove irrelevant functional variables, but also to retain the most relevant wavelet-based features in stress level recognition.

5.1 Step 1. Wavelet decomposition of the physiological functional variables

Recall that five physiological signals are considered, namely *Hand EDA*, *Foot EDA*, *HR*, *RESP* and *EMG*. In a first stage, we perform a functional variable decomposition using the Haar wavelet, which is considered as the simplest one. This choice is sufficient for our analysis since the main goal at this stage is to benefit from a basic scale information without the need for choosing a suitable wavelet type, which is required for instance, in the detection of some specific events or patterns. We recall that the Haar wavelet's mother wavelet function $\psi(t)$ is given by:

$$\psi(t) = \begin{cases} 1 & 0 \leq t < \frac{1}{2}, \\ -1 & \frac{1}{2} \leq t < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

and the corresponding scaling function $\phi(t)$ is given by:

$$\phi(t) = \begin{cases} 1 & 0 \leq t < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

We have opted for a full wavelet decomposition corresponding to 12 decomposition levels, which correspond to the maximum levels compatible with $2^{12} = 4096$ samples. For all computational tasks, we used the R software (R Core Team 2016), with the randomForest package proposed by Breiman and Cutler (2015) and RFgroove package developed by Gregorutti (2016).

5.2 Step 2. Physiological functional variables removal

In this section, the importance of physiological functional variables considered in our study will be presented. Since not all of these variables are important, we will proceed
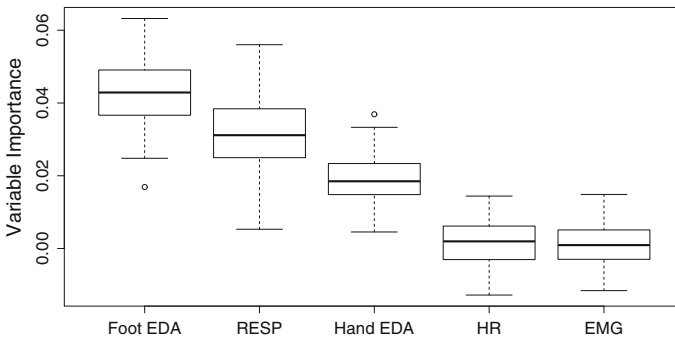
Fig. 3 Boxplots of grouped VI by physiological signals for 100 executions

to the elimination of the less important physiological signals, which do not contribute significantly in the identification of stress level.

Importance of physiological variables

Data resulted from the physiological functional variable decomposition provide $4096\times$ 5 scalar variables corresponding to the wavelet coefficients. Taking advantage of the grouped VI, the scalar variables were grouped by physiological variable and the VI of each considered group was computed. Thereby, the grouped VI allowed to consider only 5 informative VI instead of 20,480 VI of the scalar variables.

In order to examine the VI of each group, we present in Fig. 3 the boxplots of the grouped VI values, computed for 100 executions.

It can be noticed that *Foot EDA* is the most important physiological variable with a VI distribution around 4%. *RESP* is the second most important variable since the median value of the distribution of 100 VI is about 3%. *Hand EDA* comes out important when compared to the *EMG* and *HR*. The distribution of its 100 VI is around 2%. *Foot EDA* and *Hand EDA* are found to be among the most important variables. This result is in agreement with Healey and Picard (2005) findings that EDA is among the most correlated variable with the driver's stress level. *EMG* and *HR* variables are found to have a distribution of values corresponding to 100 VI around 0.

RF-RFE on the grouped physiological functional variables

In order to eliminate the less relevant physiological variables, we applied the RF-RFE algorithm ten times to the wavelet coefficients grouped by functional variables. Each execution offers a ranking of the 5 variables and provides the list of variables retained by the selected procedure. Table 6 shows the result of 10 executions of the RF-RFE algorithm on the five physiological variables. The italicized values correspond to the final retained variables. For instance, *Foot EDA*, *RESP*, and *Hand EDA* are selected in the first execution.

Based on these 10 executions, *Foot EDA* is always included in the selected model. Additionally, *EMG* and *HR* (except for one execution) are excluded from the list

Table 6 Selected model for 10 executions of the RF-RFE algorithm

Ex. Num	Weights					Num sel var
	5	4	3	2	1	
1	*Foot EDA*	*RESP*	*Hand EDA*	HR	EMG	3
2	*HR*	*RESP*	*Hand EDA*	*Foot EDA*	EMG	4
3	*Foot EDA*	Hand EDA	HR	EMG	RESP	1
4	*Foot EDA*	RESP	Hand EDA	EMG	HR	1
5	*RESP*	*Foot EDA*	Hand EDA	HR	EMG	2
6	*Foot EDA*	*RESP*	Hand EDA	EMG	HR	2
7	*Foot EDA*	Hand EDA	RESP	HR	EMG	1
8	*Foot EDA*	RESP	Hand EDA	HR	EMG	1
9	*Foot EDA*	Hand EDA	RESP	EMG	HR	1
10	*Foot EDA*	*RESP*	*Hand EDA*	HR	EMG	3

The italicized values correspond to the retained variables

Fig. 4 Endurance score of the five physiological variables for 100 runs. The two last physiological variables are removed
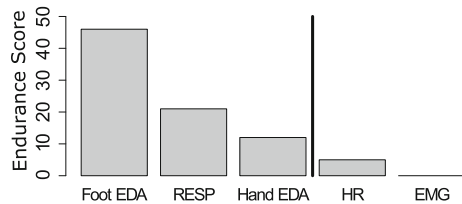
of the variables selected in the model. It can be noticed that the number of selected variables and the order of importance of these selected variables vary from an iteration to another. Thus, we propose to use the procedure detailed in Table 4 in order to aggregate the information contained in these 10 executions and to select the most "enduring" physiological variables. Figure 4 displays the "endurance" score of the five physiological variables computed based on equation (7). The results confirm the order found by the VI measure (see Fig. 3) where both *EMG* and *HR* had low endurance score. Moreover, the endurance score reveals that *Foot EDA* is the most important variable in driver's stress level classification, followed by *RESP* and *Hand EDA*. Thus it is confirmed again that the EDA is an important variable in driver's stress level recognition. These findings corroborate with the findings of Healey and Picard (2005), who reported that features extracted from the *EDA* and *HR* have the highest correlation with the stress metric built from the driving video recordings. However, it should be noted that the *Foot EDA* was not considered in the original study. The original study (Healey and Picard 2005) determined the most correlated physiological signals to driver's stress using features extracted from four physiological variables, namely the mean and the standard deviation of the *Hand EDA*, *EMG*, *RESP*, the mean and 4 metrics extracted from the *HR*.

In contrast with the results of Healey and Picard (2005), *RESP* is selected in our analysis. This can be explained by the fact that the wavelet-based features used here

may be more informative than the mean and the standard deviation used in the original study. In addition, the *RESP* wavelet features are proved sensitive in detecting the transient changes in respiration signal (Alkali et al. 2014). *HR* was found to be of less relevance in our proposed model. This may be due to additional information contained in the expert-based features that are not present in the proposed wavelet features.

Our analysis having included both Hand and Foot EDA, has revealed that the Foot EDA is as important as Hand EDA and may be even more relevant for the experienced stress levels of the drivers. Usually, EDA is measured on the fingers or on the palms. However, these two placements are not preferred in real-world task performance because they may hinder hand functionality. Measuring EDA on foot or shoulders is more practical, while having no significant difference from the signal captured from the fingers (Van Dooren et al. 2012). Several other studies have also shown the importance of the placement of the EDA sensor (Picard et al. 2016; Boucsein 2012; Van Dooren et al. 2012). Hence, one can suggest to consider Foot EDA in future studies that aim to recognize stress levels experienced in real-world driving.

Based on the "endurance" score, the final model retains the three physiological variables: *Foot EDA*, *RESP*, and *Hand EDA* as most relevant in stress level classification. In the next subsection, the wavelet levels of these selected three variables will be considered. We will then seek the model based on their most "enduring" wavelet levels to the selection procedure.

5.3 Step 3. Wavelet levels selection from the three retained physiological variables

The goal of this step is to identify, for the retained physiological functional variables, the corresponding wavelet levels most able to predict the driver's stress class. To achieve this goal, the wavelet coefficients are now grouped by levels.

Let us denote the approximation of the signal s as $A12$ and the detail at level j by Dj then the wavelet decomposition can be written as follows

$$\tilde{s} = A12 + \sum_{j=1}^{12} Dj, \tag{14}$$

where $A12(t)$ is reconstructed from the single scaling coefficient at level 12 and details of increasing resolution Dj are reconstructed from the DWT wavelet coefficients $(d_{j,k})$ [see Eq. (11)].

Wavelet levels importance

Let V1 stand for *Hand EDA*, V2 for *Foot EDA*, and V5 for *RESP*. We simply denote by Vx_A12 or Vx_Dj the wavelet decomposition components of the functional variable Vx.
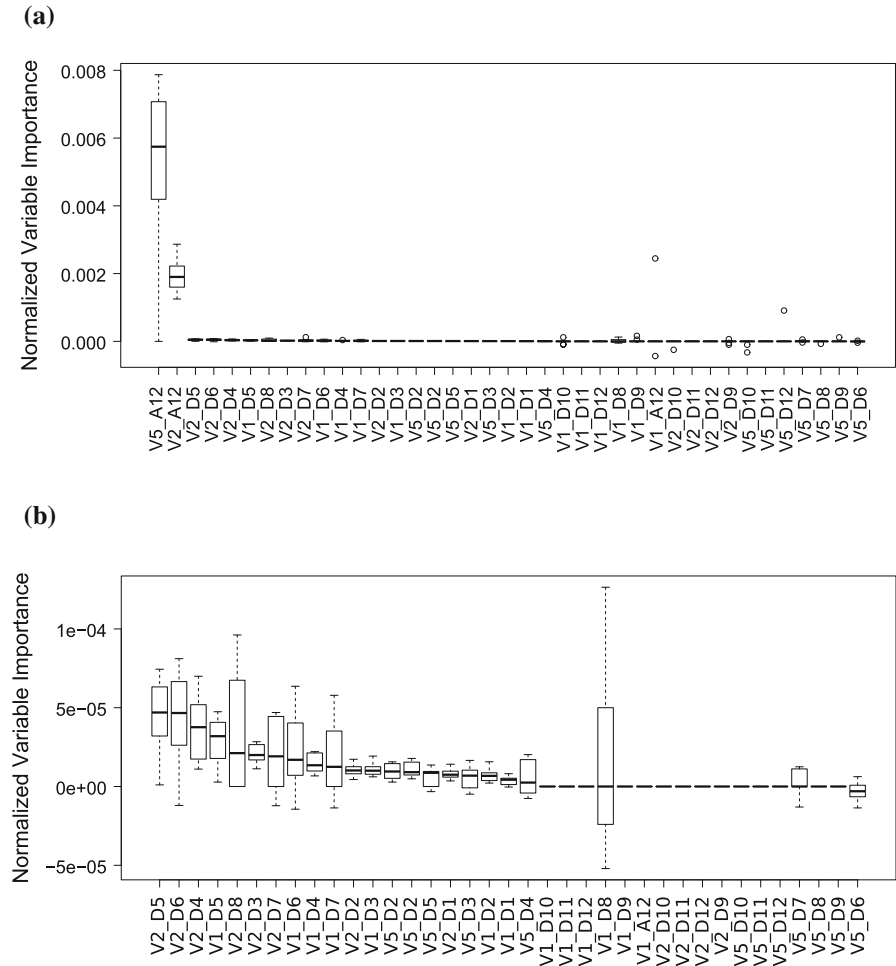
(a)

(b)

Fig. 5 Grouped VI of the wavelet levels for 10 iterations. **a** Grouped VI of the V1, V2 and V5 wavelet levels, **b** grouped VI of the V1, V2 and V5 wavelet levels without V2_A12 and V5_A12

Figure 5a depicts the distribution of the VI computed over 10 executions of the different wavelet levels of the three physiological variables. The distributions show that the approximation level ($A12$) of both *Foot EDA* and *RESP* are the most important wavelet levels in the stress level classification. The variation of the distribution of the other wavelet levels is not clearly visible in Fig. 5a. In order to better investigate the relative variation of small VI, we present a graph obtained by removing the dominant variables $V2\_A12$ and $V5\_A12$ and the outliers in Fig. 5b. We find that for the *Hand EDA* (V1), levels from 1 to 8 are the most important. Wavelet levels from 3 to 8 are important for the *Foot EDA* (V2). For the *RESP* (V5), levels 1 and 2 are the most important.

In the next paragraph, we compute the "endurance" score for all wavelet levels that allows to select the final model.
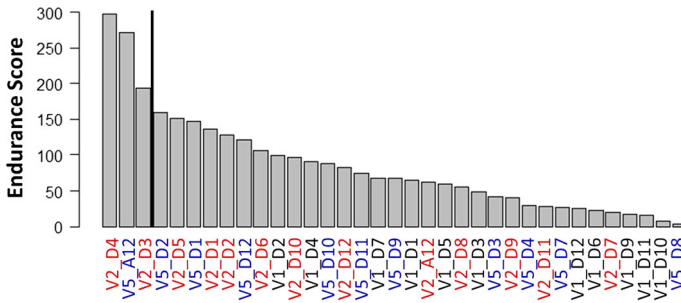
Fig. 6 Wavelet levels selection for 10 executions

Wavelet levels selection

When applying the RF-RFE algorithm to the groups composed by the wavelet levels
of the three retained variables several times, the selected model varies in terms of the
number of the selected levels, the validation error and even the ranking of the different
wavelet levels. In order to reduce this variability, an "endurance" score, introduced
earlier by Equation (8), is proposed. It allows to rank the different wavelet levels
according to their ability to persist in the selection procedure. The plot of the ranked
variables using this score is displayed in Fig. 6. There are two ways to select variables
of the final model, when examining the consecutive differences of the "endurance"
scores: one can start from the highest value and cut off at the largest gap of this
difference or start from the lowest score and cut off at the largest gap of this difference.
When applying the first approach on the wavelet details corresponding to the three
retained physiological variables, we select $V2\_D4$ and $V5\_A12$. If the second way
of selection is chosen, $V2\_D4$, $V5\_A12$, and $V2\_D3$ will be selected. Three wavelet
signals that correspond to levels 4 and 3 of details and the approximation at level 12,
among $39 = 13 \times 3$ wavelet signals, are selected in our case. This can be justified by
the fact that the third level ($V2\_D3$) is not too different from $V2\_D4$.

Based on the selected wavelet levels, we examine then the reconstructed physio-
logical signals. For instance, Foot EDA wavelet details on levels 3 and 4 and RESP
wavelet approximation at level 12 corresponding to Drive 07 are depicted in Fig. 7.
Indeed, using only two selected levels for Foot EDA allows to distinguish between the
three segments of city, highway, and rest. The reconstructed approximation level of
the respiration (Fig. 7) highlights the fact that the 3 periods can be easily recognized.
It can be also noticed that the value of RESP approximation level decreases when the
stress level increases which is aligned with the conclusion of the study of Lin et al.
(2011), showing that stress is negatively correlated with the respiration rate.

An additional "unidimensional" information

In this subsection, we propose to investigate, for each physiological functional vari-
able, wavelet levels that are the most able to classify drivers stress level. Therefore,
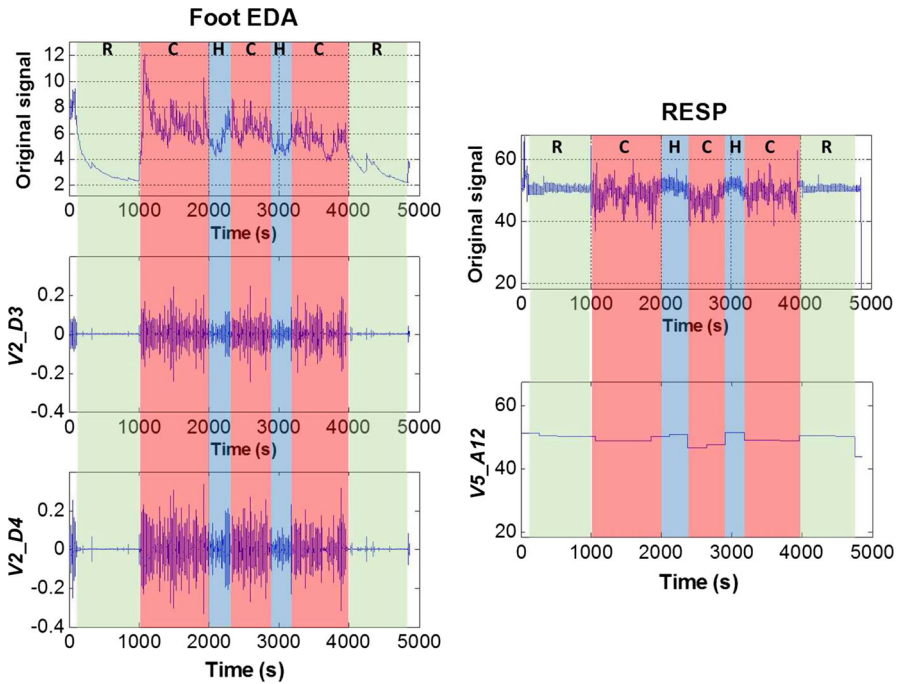three models, aiming each at recognizing stress level based on a given physiological

Fig. 7 Illustration of the reconstructed signals corresponding to Drive 07 for Foot EDA (left column) and RESP (right column), based on the three selected wavelet levels (see Fig. 6). The letter "R" corresponds to rest period, "C" to city driving and "H" to highway driving

variable, are considered. The selection of wavelet levels is then performed independently, for model composed by *Hand EDA*, *Foot EDA* and *RESP* respectively. For that, Equation (8) is used first to compute the "endurance" score of the different wavelet levels of the *Foot EDA*. We note that two wavelet detail levels are selected, based on the "endurance" score shown in Fig. 8a, which are namely levels 4 and 5. For this physiological predictor, the wavelet level 4 was selected in the final model that combines the different wavelet levels of the three physiological predictors.

Figure 8b reveals that the first and the second details of the wavelet levels and the approximation at level 12 are selected for the RESP predictor.

Figure 8c shows *Hand EDA* wavelet levels ordered by endurance score. When starting from the smallest value of the score and looking at the consecutive difference, levels 1 to 8 (except 7) are selected for *Hand EDA*. The same result was found based on VI analysis of *Hand EDA* levels (see Fig. 5b). In addition, the "endurance" score proposes that $D4$, $D5$ then $D6$, are in the same order as for *Foot EDA* (see Fig. 8c for comparison).

When considering separate models composed by each physiological variable, the proposed endurance score provides a similar order of the wavelet levels as given by VI of the global model (see Fig. 5). This additional separate analysis supports the findings based on the proposed endurance score.

In the next subsection, the performance of the 3-step based approach will be presented and an assessment of the final misclassification rate will be performed.
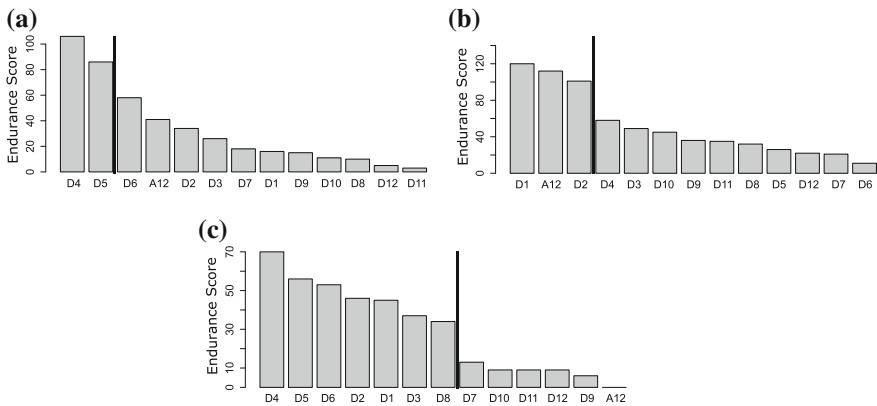
Fig. 8 Wavelet levels endurance score of the three retained physiological variables after 10 executions. **a** Foot EDA, **b** RESP, **c** Hand EDA

5.4 Performances of the proposed "blind" approach

Recall that the "blind" 3-step approach applied on the *drivedb* database resulted in the selection of Hand EDA, Foot EDA and RESP as the most relevant functional variables, in addition to their most important wavelet decomposition levels. Specifically, this resulted in a total of 3 wavelet signals corresponding to levels 3 and 4 of the Foot EDA wavelet details and level 12 of the RESP wavelet approximation.

In this section, we first assess the performances of the RF-based model when considering these retained variables in the classification of the driver's stress level, in terms of misclassification rate. Then we perform a cross-validation like-procedure for model error estimation taking into account the particularity of the cohort used in the experiments. Finally, we proceed with evaluation of the resulting performance of the proposed "blind" approach, when compared to those obtained by the "expert-based" approach.

Recall that a total of 669 corresponding wavelet coefficients were finally considered as input variables to the RF algorithm (for the *Foot EDA*, $512 = 2^{(12-3)}$ and $256 = 2^{(12-4)}$ coefficients and 1 coefficient corresponding to the level 12 of the *RESP* wavelet approximation).

In order to evaluate the performances of the proposed blind approach, two different procedures were considered: the first one consists in evaluating the error of the model, composed of the whole set of variables. The second procedure considers the error resulting from a wavelet level selection among the three signals corresponding to the retained wavelet levels using the RF-RFE method summarized in Table 3.

The model error corresponds to the average of the misclassification rate obtained in each iteration of the evaluation procedure, which is repeated 100 times in order to reduce the variability of the estimation. The misclassification rate is estimated in two different ways: the error obtained using a test set ($\frac{1}{3}$ of the data) and the OOB error embedded in the RF algorithm.

As can be seen in Table 7a, the model errors resulting from the wavelet selection are lower than those obtained when considering the complete set of coefficients. The model

Table 7 Model error: misclassification rate averaged over 100 executions

(a) Performances of the "blind" 3-step approach

Input variables	Test error	OOB error
With selection	0.28 (sd = 0.09)	0.27 (sd = 0.02)
Without selection	0.31 (sd = 0.09)	0.31 (sd = 0.03)

(b) Performances of the "blind" 3-step and expert-based approaches for cross validation like procedure. Config1 (training set size=54 and test set size=14), Config2 (training set size=48 and test set size=20) and Config3 (training set size=40 and test set size=28)

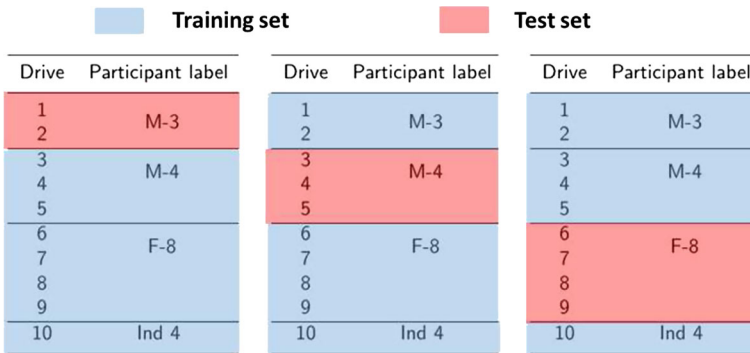| Input variables	Selection	Config1	Config2	Config3	Mean
Wavelet features	With selection	0.19 (sd=0.05)	0.22 (sd=0.04)	0.26 (sd=0.03)	0.22
	Without selection	0.31 (sd=0.04)	0.22 (sd=0.05)	0.27 (sd=0.04)	0.27
Expert-based features		0.33 (sd=0.01)	0.14 (sd=0.02)	0.25 (sd=0.04)	0.25

Fig. 9 Three configurations of training and test sets choice in the cross validation like procedure

error assessed on the test set with wavelet level selection is 0.28 ($sd = 0.09$) while the OOB error is 0.27 ($sd = 0.02$). When the 669 coefficients are considered, the OOB error is 0.31 ($sd = 0.03$) and the error estimated using the test set is 0.31 ($sd = 0.09$).

Recall that the *drivedb* datasets are based on a real-world driving of four participants, some of whom repeated the experiment. In order to compute the misclassification rate, we adopt a particular cross-validation procedure taking into account this particularity of the experiments (see Fig. 9).

In fact, the error assessment was evaluated for two models: one composed of the whole set of features (without selection) and the second procedure considered the resulting error based on a wavelet level selection, among the retained three levels using RF-RFE (with selection).

In Table 7b, we consider the three configurations of the cohort disposition defined in Fig. 9. The partitioning of the dataset into 3 configurations is done as follows: Config k consists in considering the data of the driver k, $k = 1, \ldots, 3$ as the test set, while the retained data related to the other drivers is considered as the training set as shown in Fig. 9. We note that the data related to driver 4 was considered in the training set

Table 8 Features related to the EDA (Foot and Hand) and to the respiration signals proposed by Healey and Picard (2005)

Type of features	Signals	Features
6 Statistical features	Hand EDA, Foot EDA and RESP	Mean
		Standard deviation
4 Power spectral power features	RESP	Spectral power density in the bands 0–0.4 Hz
8 Startle features	Hand EDA and Foot EDA	Total number of startle
		Sum of the magnitudes
		Sum of the duration
		Sum of the areas under the startles

of the 3 configurations since he performed only one test drive. Table 7b shows the mean error over the 3 configurations of the "blind" approach in both cases: with and without selection, in addition to the model error resulting from the RF using the 18 expert-based features shown in Table 8. The estimated model error considering the wavelet features (without selection) is 0.27 while it is 0.22 with selection. We note that the expert-based approach had a model error of 0.25.

The expert-based features, proposed by Healey and Picard (2005), was also used in the Deng et al. (2012) study, which relied on the *drivedb* dataset as well. In such study, the PCA method was used for variable selection. The misclassification rate was estimated using ten-fold cross validation approach. The misclassification rate average based on five fusion algorithms (including LDF, C4.5, SVM, NB and KNN), applied to the PCA selected features, was found to be equal to 0.30 ($sd = 0.07$) while the misclassification rate of the expert based features (22 features) is 0.33 ($sd = 0.10$). The best misclassification rate, estimated using the PCA selected features and the SVM as fusion method, was found to be equal to 0.21 ($sd = 0.03$). The worst misclassification rate, estimated considering the whole features with KNN was equal to 0.38 ($sd = 0.07$). We note that when considering the LDF method, used as well in Healey and Picard (2005), the misclassification rate was found to be around 0.34 for the two sets of features (complete set of features and PCA selected features).

In summary, we note that our blind approach performances are comparable (within the same range) of the expert based ones.

6 Discussion

In this section, we summarize the rationale for our approach and discuss the obtained results. First, we recall that this study was based on the open PhysioNet *drivedb* database, which is a subset of the original data constructed and used in Healey and Picard (2005) study.

Most studies using *drivedb* data relied on features extraction as a basis for drivers stress level classification in real-world driving experiments. This is in contrast with our study where a functional variable approach, coupled with a multi-resolution wavelet

decomposition, are considered as the input for Random Forest-based classification procedure.

The developed approach consisted, at a first stage, in functional variables decomposition using the Haar wavelet. The irrelevant functional variables were removed based on a proposed endurance score. This score reflected the variable ability to persist (in the sense of being selected and well ranked) when applying the variable selection procedure 10 times, based on RF-RFE and grouped variable importance. The corresponding wavelet levels of the retained functional variables were ranked based on the same score. The most "enduring" wavelet levels to the selection procedure were finally retained.

The developed procedure, inspired from Gregorutti et al. (2015) work, was applied considering only models composed of p, $p-1$, ..., 1 variables, where p is the number of variables. Thus, not all variable combinations were considered, which could lead to lower error rate. In fact, when p is large it is computationally costly to consider all variable combinations.

Recall that our "blind" 3-step approach allowing to classify driver's stress level using various physiological measurements, resulted in the ordered selection of the *Foot EDA*, *RESP* and *Hand EDA*. Moreover, a finer wavelet-based analysis suggested a total of 3 wavelet signals corresponding to the levels 3 and 4 of *Foot EDA* wavelet details and the approximation corresponding to level 12 of *RESP*. The fact that the EDA emerges as important in driver's stress level recognition is expected since Healey and Picard (2005) original study reported that features related to *EDA* and *HR* are the most correlated to the stress metric extracted from video-based subjective evaluations.

Furthermore, our findings are in alignment with more recent studies (Picard et al. 2016; Boucsein 2012), which revealed the importance of the EDA sensor placement in a given experiment. In that regard, our study, which included both Hand and Foot EDA, has indicated that the Foot EDA is as important and may be even more relevant in detecting stress levels of the drivers in real-world driving situations.

Several performance indexes were evaluated using different error rates. For that, we assessed the model performances, in terms of misclassification rate, with regards to the retained variables in the classification of the driver's stress level. We then performed a cross-validation like-procedure for model error estimation, taking into account the particularity of the cohort used in the experiments. Finally, we examined the performances of our model using the proposed "blind" approach, as it compares with the results from the "expert-based" approach. The results show that our blind approach performances were comparable (i.e. within the same range) to the expert based ones.

7 Conclusion

This study is based on physiological functional variable selection for driver's stress level classification using random forests. The physiological datasets used here consisted in ten real-world driving experiments extracted from the open *drivedb* database. Several physiological signals were collected using portable sensors during real-world driving experiments in three types of routes (rest area, city, and highway) in the Boston

area. These signals are: the electrodermal activity measured on the driver's left hand and foot, electromyogram, respiration, and the heart rate.

The contributions of this work are twofold since they touched on the method as well as the application aspects of this problem. The proposed approach was based on RF Recursive Feature Elimination (RF-RFE) and grouped variable importance applied to two different levels of data selection strategies: physiological and wavelet-based variables. For that, all physiological signals were decomposed on a common Haar wavelet basis and then analyzed in search of important variables using a proposed "endurance" score.

The developed method provided a "blind" procedure (without prior information) of driver's stress level classification that resulted in a model error close to the expert-based approach performance. In addition, our approach suggested new physiological features based on the wavelet levels. Moreover, the proposed approach offered a ranking of physiological variables according to their importance in drivers stress level classification in city versus highway driving, with restful period for baseline reference. For the considered case of study, the analysis results suggest that the electromyogram and the heart rate signals are less relevant when compared to the electrodermal and the respiration signals. Furthermore, the electrodermal activity measured on the driver's foot emerged as more relevant than that captured on the hand.

The different changes of road type are the same for all the experiments, since all the drivers follow the same road-map, but occur at different instants. By using classical time warping techniques, one could synchronize these events across the experiments. This would enable the interpretation of each wavelet coefficient and could allow a deeper analysis of the signal by adding an other step to select the most important coefficients within the details selected in Step 3 of our procedure.

One way to further improve the methodology adopted in this study would be to consider costs that penalize the high versus low confusion. In fact, confusing a low stress level with a high one is more serious than predicting the high level as medium one, or the medium as high stress level. This can be easily achieved by introducing different misclassification costs to grow each tree of the forest. The second variant is a preprocessing of the segments extracted from the driving experience recordings. A pre-selection of the coefficients in order to reduce dimension can be performed using wavelet based techniques or Principal Components Analysis for instance, to ensure better stability of the variable selection process. Finally, we could also consider a preprocessing of each driving experience considered as a whole and redefine new original signals realigned on a single time interval by synchronizing typical events.

In this study, the variable of interest was built according to the type of route traveled by the driver, thus it depended mainly on the hypothesis assuming that the stress level increases when driving in the city and decreases when the participant is at the rest. Such hypothesis remains simplified since it does not take into account several factors such as the driver's cognitive state, mental workload, and anticipation of some situations. The effect of the external visual and sonic environmental complexity could as well affect the driving performance (Horberry et al. 2006) thus the driver's stress level. Additional information can be gathered using measurements characterizing the affective state of the driver, the driving environment and the inside vehicle complexity

(see El Haouij et al. 2018). Thereby, a finer analysis of the groupings can be proposed for the coefficients resulting from wavelet decompositions as well.

References

Akbas A (2011) Evaluation of the physiological data indicating the dynamic stress level of drivers. Sci Res Essays 6(2):430–439

Alkali AH, Saatchi R, Elphick H, Burke D (2014) Short-time Fourier and wavelet transform analysis of respiration signal obtained by thermal imaging. In: 2014 9th International Symposium on Communication Systems, Networks & Digital Sign (CSNDSP). IEEE, pp 183–187. https://doi.org/10.1109/CSNDSP.2014.6923821

Auret L, Aldrich C (2011) Empirical comparison of tree ensemble variable importance measures. Chemometr Intell Lab Syst 105(2):157–170. https://doi.org/10.1016/j.chemolab.2010.12.004

Ayata D, Yaslan Y, Kamasak M (2016) Emotion recognition via random forest and galvanic skin response: comparison of time based feature sets, window sizes and wavelet approaches. In: 2016 Medical Technologies National Congress (TIPTEKNO). IEEE, pp 1–4. https://doi.org/10.1109/TIPTEKNO.2016.7863130

Bach FR (2008) Consistency of the group lasso and multiple kernel learning. J Mach Learn Res 9(Jun):1179–1225

Bostrom J (2005) Emotion-sensing PCs could feel your stress. PC World

Boucsein W (2012) Electrodermal activity. Springer, Berlin

Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. The wadsworth and Brooks–Cole statistics-probability series. Taylor & Francis, London

Breiman L, Cutler A (2015) Randomforest: Breiman and cutler's random forests for classification and regression. R Package Version 46-12 http://cran.r-project.org/package=randomForest

Chaudhary R (2013) Electrocardiogram comparison of stress recognition in automobile drivers on matlab. Adv Electron Electr Eng 3(8):1007–1012

Deng Y, Wu Z, Chu C, Yang T (2012) Evaluating feature selection for stress identification. In: Information Reuse and Integration (IRI), 2012 IEEE 13th international conference on, pp 584–591. https://doi.org/10.1109/IRI.2012.6303062

Díaz-Uriarte R, de Andrés SA (2006) Gene selection and classification of microarray data using random forest. BMC Bioinform 7(1):1–13. https://doi.org/10.1186/1471-2105-7-3

El Haouij N, Poggi JM, Sevestre-Ghalila S, Ghozi R, Jaïdane M (2018) AffectiveROAD system and database to assess driver's attention. In: SAC 2018: symposium on applied computing, April 9–13, Pau. https://doi.org/10.1145/3167132.3167395

Ferraty F, Vieu P (2006) Nonparametric functional data analysis: theory and practice (springer series in statistics). Springer-Verlag New York Inc., Secaucus

Genuer R, Poggi JM, Tuleau-Malot C (2010) Variable selection using random forests. Pattern Recognit Lett 31(14):2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014

Genuer R, Poggi JM, Tuleau-Malot C (2015) VSURF: an R package for variable selection using random forests. R J 7(2):19–33

Goldberger A, Amaral L, Glass L, Hausdorff J, Ivanov P, Mark R, Mietus J, Moody G, Peng CK, Stanley H (2000) Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation 101(23):e215–e220

Granero AC, Fuentes-Hurtado F, Naranjo Ornedo V, Guixeres Provinciale J, Ausín JM, Alcañiz Raya M (2016) a Comparison of physiological signal analysis techniques and classifiers for automatic emotional evaluation of audiovisual contents. Front Comput Neurosci 10:74. https://doi.org/10.3389/fncom.2016.00074

Gregorutti B (2016) RFgroove: importance measure and selection for groups of variables with random forests. R Package Version 11 http://cran.r-project.org/package=RFgroove

Gregorutti B, Michel B, Saint-Pierre P (2015) Grouped variable importance with random forests and application to multiple functional data analysis. Comput Stat Data Anal 90:15–35. https://doi.org/10.1016/j.csda.2015.04.002

Gregorutti B, Michel B, Saint-Pierre P (2016) Correlation and variable importance in random forests. Stat Comput. https://doi.org/10.1007/s11222-016-9646-1

Guendil Z, Lachiri Z, Maaoui C, Pruski A (2015) Emotion recognition from physiological signals using fusion of wavelet based features. In: 2015 7th International Conference on Modelling, Identification and Control (ICMIC), IEEE, pp 1–6. https://doi.org/10.1109/ICMIC.2015.7409485

Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46(1–3):389–422. https://doi.org/10.1023/A:1012487302797

Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Springer series in statistics. Springer New York Inc., New York

Healey JA (2000) Wearable and automotive systems for affect recognition from physiology. Ph.D. Thesis, MIT Department of Electrical Engineering and Computer Science

Healey JA, Picard RW (2005) Detecting stress during real-world driving tasks using physiological sensors. IEEE Trans Intell Transp Syst 6(2):156–166. https://doi.org/10.1109/TITS.2005.848368

Horberry T, Anderson J, Regan MA, Triggs TJ, Brown J (2006) Driver distraction: the effects of concurrent in-vehicle tasks, road environment complexity and age on driving performance. Accid Anal Prev 38(1):185–191

Imam MH, Karmakar CK, Khandoker AH, Palaniswami M (2014) Effect of ECG-derived respiration (EDR) on modeling ventricular repolarization dynamics in different physiological and psychological conditions. Med Biol Eng Comput 52(10):851–860

Jolliffe I (2012) Principal Component Analysis. Springer, Berlin

Karmakar C, Imam MH, Khandoker A, Palaniswami M (2014) Influence of psychological stress on QT interval. Computing in cardiology 2014:1009–1012

Lin HP, Lin HY, Lin WL, Huang ACW (2011) Effects of stress, depression, and their interaction on heart rate, skin conductance, finger temperature, and respiratory rate: sympathetic-parasympathetic hypothesis of stress and depression. J Clin Psychol 67(10):1080–1091. https://doi.org/10.1002/jclp.20833

Louppe G, Wehenkel L, Sutera A, Geurts P (2013) Understanding variable importances in forests of randomized trees

Lykken DT (1972) Range correction applied to heart rate and to GSR data. Psychophysiology 9(3):373–379. https://doi.org/10.1111/j.1469-8986.1972.tb03222.x

Mallat SG (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Mach Intell 11(7):674–693

Nicodemus KK, Malley JD, Strobl C, Ziegler A (2010) The behaviour of random forest permutation-based variable importance measures under predictor correlation. BMC Bioinform 11(1):1–13. https://doi.org/10.1186/1471-2105-11-110

Picard RW, Fedor S, Ayzenberg Y (2016) Multiple arousal theory and daily-life electrodermal activity asymmetry. Emot Rev 8(1):62–75. https://doi.org/10.1177/1754073914565517

Poggi JM, Tuleau C (2007) Classification of objectivization data using cart and wavelets. In: Proceedings of the IASC 07. Aveiro, pp 1–8

R Core Team (2016) R: A language and environment for statistical computing. In: R foundation for statistical computing. Vienna. www.r-project.org

Ramsay JO, Silverman BW (2002) Applied functional data analysis: methods and case studies, vol 77. Springer, New York

Ramsay JO, Silverman BW (2005) Functional data analysis. Springer, New York. https://doi.org/10.1007/b98888

Rigas G, Katsis C, Bougia P, Fotiadis D (2008) A reasoning-based framework for car drivers stress prediction. In: Control and automation, 2008 16th mediterranean conference on. pp 627–632. https://doi.org/10.1109/MED.2008.4602162

Sharma N, Gedeon T (2012) Objective measures, sensors and computational techniques for stress recognition and classification: a survey. Comput Methods Programs Biomed 108(3):1287–301. https://doi.org/10.1016/j.cmpb.2012.07.003

Sidek KA, Khalil I (2011) Automobile driver recognition under different physiological conditions using the electrocardiogram. PC World 38:753–756

Singh RR, Conjeti S, Banerjee R (2012) Biosignal based on-road stress monitoring for automotive drivers. In: 2012 National Conference on Communications (NCC), IEEE, pp 1–5. https://doi.org/10.1109/NCC.2012.6176845

Singh M, Queyam AB (2013) Stress detection in automobile drivers using physiological parameters: a review. Int J Electron Eng 5(2):1–5

Smart RG, Cannon E, Howard A, Frise P, Mann RE (2005) Can we design cars to prevent road rage? Int J Veh Inf Commun Syst 1(1–2):44–55. https://doi.org/10.1504/IJVICS.2005.007585

Strobl C, Zeileis A (2008) Danger: high power!? exploring the statistical properties of a test for random forest variable importance. In: Proceedings of 18th international conference on computational statistics

Tao J, Tan T (2005) Affective computing: a review. In: International conference on affective computing and intelligent interaction. Springer, pp 981–995

Ullah S, Finch CF (2013) Applications of functional data analysis: a systematic review. BMC Med Res Methodol 13(1):43

Van Dooren M, De Vries JJ, Janssen JH (2012) Emotional sweating across the body: Comparing 16 different skin conductance measurement locations. Physiol Behav 106(2):298–304. https://doi.org/10.1016/j.physbeh.2012.01.020

Verikas A, Gelzinis A, Bacauskiene M (2011) Mining data with random forests: a survey and results of new tests. Pattern Recognit 44(2):330–349. https://doi.org/10.1016/j.patcog.2010.08.011

Yang K, Yoon H, Shahabi C (2005) A supervised feature subset selection technique for multivariate time series. In: Proceedings of the workshop on feature selection for data mining: interfacing machine learning with statistics, pp 92–101

Zhang L, Tamminedi T, Ganguli A, Yosiphon G, Yadegar J (2010) Hierarchical multiple sensor fusion using structurally learned Bayesian network. In: Wireless health 2010 on—WH '10. ACM Press, New York, p 174. https://doi.org/10.1145/1921081.1921102

Zhu R, Zeng D, Kosorok MR (2012) Reinforcement learning trees. Technical reports on University of North Carolina