Identification of driving simulator sessions of depressed drivers: A comparison between aggregated and time-series classification

Dr. Christos Katrakazas*
Post-Doctoral Research Associate
National Technical University of Athens
Department of Transportation Planning and Engineering
5 Iroon Polytechniou St., GR-15773, Athens, Greece
Email: ckatrakazas@mail.ntua.gr

Prof. Constantinos Antoniou
Professor
Chair of Transportation Systems Engineering
Department of Civil, Geo and Environmental Engineering
Technical University of Munich, Munich, 80333
Email: c.antoniou@tum.de

Prof. George Yannis
Professor
National Technical University of Athens
Department of Transportation Planning and Engineering
5 Iroon Polytechniou St., GR-15773, Athens, Greece
Email: geyannis@central.ntua.gr

Abstract

Depression has been found to significantly increase the probability of risky driving and involvement in traffic collisions. The majority of studies correlating depressive symptoms with driving, pursue to predict the differences in driving behavior if the driver has already been diagnosed. Little evidence can be found, however, on how mental and psychological disorders can be identified from driving data, and usually analyses utilize simple models and aggregated data. This study aims at utilizing microscopic data from a driving simulator to detect sessions belonging to “depressed” drivers by utilizing powerful machine learning classifiers. Driving simulator sessions from 11 older drivers with symptoms of depression and 65 healthy drivers were utilized towards that aim. Random Forests, an ensemble classifier, with proven efficiency among transportation applications, are then trained on highly disaggregated data describing the mean and standard deviation of speed and lateral or longitudinal acceleration of drivers in the simulator. The kinematic data were aggregated in 30-seconds, 1-minute and 5-minute intervals, but the corresponding time-series of the measurements were also taken into account. Furthermore, classifiers were treated with imbalanced learning techniques to address the scarcity of depressed drivers among the healthy. Time-series of mean speed and the standard deviation of longitudinal acceleration even with a duration of 30-seconds have proven to be the best predictors of driving sessions belonging to depressed drivers with a very low rate of false alarms. The results outperform previous approaches, and indicate that naturalistic driving data or deep learning could prove even more efficient in detecting depression.

Keywords: Depression, Driving Simulator, Random Forests, Time-Series

1. Introduction

Depression is one of the two most common mental disorders among the world's population (the other one being anxiety), influencing more than 300 million people in 2017 (World Health Organization, 2017). Furthermore, the total number of people suffering from depressive symptoms, has significantly increased during the last 15 years (World Health Organization, 2017). Usually, depressive disorders can be divided into two categories: i) major depressive disorders (also referred to as episodes), where symptoms include a depressed mood, loss of interest and decrease of energy for a short amount of time and ii) dysthymia, where the aforementioned symptoms might be milder but occur on a chronic basis. The significance of depression for mental well-being can be linked to the fact that is prevalent among different age groups, e.g. adolescents and young adults (Mojtabai et al., 2016) as well as elderly (Fiske et al., 2012) and is also heavily correlated with mild cognitive impairment (MCI; Ballenger, 2008; Beratis et al., 2017).

Recent studies (Beratis et al., 2017; Brunnauer and Laux, 2017; Bulmash et al., 2006; Pavlou, 2016; Scott-Parker et al., 2013a) have investigated the relationship between depression and driving performance and their findings conclude to a deterioration of driving behaviour in patients with depressive symptoms. However, in the majority of recent studies, the research hypothesis is based on a pre-defined cognitive impairment and is focusing on the differences between a group of depressed individuals and a control group with regards to driving behavioural (e.g. speed, acceleration, steering wheel variations) or safety characteristics (e.g. Time-to-Collision; TTC, headway or reaction times). Moreover, in order for associations with regards to depression or cognitive impairment in general and driving behaviour to be drawn, methodologies are limited to simple hypothesis testing (McDonald et al., 2018; Scott-Parker et al., 2013b), descriptive statistics (Pavlou et al., 2016), simple regression (Beratis et al., 2017), with the most sophisticated technique being Structural Equational Modelling (SEMs) (Pavlou, 2016; Scott-Parker et al., 2013b). In general, patients diagnosed with MCI, participate in driving simulator studies or on-road tests (Papadimitriou et al., 2017), and the data are post-processed and largely aggregated in order to distinguish between consistent and safe driving or unsafe behaviour.

In the current era, advances in Intelligent Transportation Systems (ITS), data collection and handling, as well as vehicular technologies have brought about new innovative concepts, such as machine learning-based driver analytics (Vlahogianni and Barmpounakis, 2017) and autonomous driving (Levinson et al., 2011). In such frameworks, the identification of driving behaviour needs to be efficient in a proactive and real-time manner, ensuring safety and comfort among all traffic participants.

Limited proof can be found in the literature on how a specific cognitive impairment, such as depression, can be identified using microscopic (i.e. highly disaggregated) driving data from experimental studies with the exploitation of machine learning approaches. This gap forms the motivation for the current paper, which aims at identifying if driving characteristics captured in a driving simulator experiment belong to a depressed driver or a control participant. The proposed approach, utilizes highly disaggregated data which describe the speed, lateral and longitudinal acceleration of participants, and compares two different methodological frameworks: i) one where data are aggregated in 30-seconds, 1-minute and 5-minute intervals and the mean and standard deviations of the aforementioned kinematic characteristics are extracted and ii) one that utilizes the time series of these kinematic variables with corresponding lengths to the aggregated data(i.e. 30-second, 1-minute and 5-minute time series). The use of different data aggregation intervals as well as the use of time series with different duration would also attempt to investigate the effectiveness of measurement duration and aggregation level in identifying data points belonging to depressed drivers, so as to determine which duration or aggregation level should be preferred by researchers and practitioners.

The remaining parts of the paper are structured as follows: initially the literature is reviewed in order to understand how depression, cognitive impairment and driving behaviour are correlated, and then the methodology to detect depression among drivers using machine learning is presented. This is followed by a description of the data utilized for the study, the results and their discussion with regards to the developed classifiers. Finally, conclusions are drawn and recommendations for further research are proposed.

2. Literature Review

Depression, as a mental impairment, affects the way in which information is processed and may lead to difficulties in rapidly changing environments which are task demanding (Grahek et al., 2019). Due to the infamy of depression as a psychiatric disorder, its potential impact on driver performance could have significant implications not only on the individual's well-being but also on traffic safety due to inabilities of depressed to concentrate and their slow reacting times (Wickens et al., 2014). Wickens et al. (2014) in their review on depression and driving, argue that although several studies have demonstrated mixed results on the effect of depression on road safety, the majority of the literature suggests an increased collision probability for depression patients. Furthermore, in the same study, it is shown that there is evidence in the literature, that depressed drivers usually drive in a more aggressive and risky way than healthy ones.

In a recent review and meta-analysis by Hill et al. (2017), it was found that a depression diagnosis, almost doubles the probability of an individual being involved in a car collision, however the large variation of study designs and study samples may hinder the transferability of results among different countries. Another important finding of Hill et al., is that the use of anti-depressant medication may lead to an enhanced collision involvement probability of 40%. As a result, driving performance of depressed drivers, might be also affected by their medications, the effects of which are usually not clear or not taken into account in studies on depression and driver behaviour. From the two aforementioned reviews (Hill et al., 2017; Wickens et al., 2014), as well as the study of Cunningham and Regan, (2016) it is demonstrated that the majority of studies, infer differences between healthy controls and depressed drivers, by comparing variables of interest, in hypothesis tests, exploratory analysis of variable variations or simple regression techniques which provide odds ratios for driving characteristics or collision involvement.

The driving performance of patients with MCI has been the study of several studies lately by Pavlou et al.(Pavlou, 2016; Pavlou et al., 2016). These studies were focused on elderly drivers with MCI, and drivers suffering from Alzheimer's or Parkinson's disease, with the aim of assessing their fitness to drive and the safety of their driving behaviour through a driving simulator experiment. It was found through exploratory analysis and the development of SEM models that MCI, Alzheimer's and Parkinson's patients had larger reaction times and increased collision probabilities than healthy controls. However, the specific effect of depression on driving behaviour and collision probability was not investigated, and the data used for the analysis were aggregated for the time duration of the driving session of each participant. Using the same dataset, Beratis et al. (2017) investigated the effect of depression on elderly drivers with MCI by using hierarchical multiple linear regression. It was again validated that depression was negatively associated with driving behaviour and that depressive symptoms can be utilized as predictors for several driving indices including speed, reaction times, lateral position and number of crashes.

The prominent characteristic of the aforementioned studies, as well as the ones included in the reviews of Hill et al., (2017) and Wickens et al., (2014), is the direction of their research hypotheses: depression or any cognitive impairments is taken as prerequisite and the effect on driving characteristics or road safety is investigated, often with simplistic approaches, in order to identify differences between patients and control cases. Little evidence can be found in the research community, of studies that are concerned with the prediction of a mental disease based on driver behaviour characteristics. To the authors knowledge, the most relevant studies are the one of Vardaki et al. (2014) and Papadimitriou et al., (2017), with the latter being the only that identified significant results. In Vardaki et al., (2014), MCI drivers and controls were attempted to be distinguished through the performance on a sign recall task in a driving simulator, but results showed that the recall performance was not able to statistically significantly predict MCI impairment. On the contrary, in Papadimitriou et al., (2017), using discriminant analysis managed to

identify almost 65% of 419 drivers, with regards to their pathological symptoms (i.e. healthy, MCI or Alzheimer's disease), based on variables captured from a driving simulator experiment. However, there were a lot of misclassification with regards to the two different cognitive disorders and the data used were aggregated for the whole duration of the driving session similarly to Pavlou, (2016).
Summarizing, it is evident from reviewing the literature on depression and cognitive impairment as well as their correlation with driving behaviour and road safety, that predicting depression symptoms from driving characteristics would be beneficial for road safety and that microscopic (highly disaggregated) data are yet to be utilized for predicting cognitive disorders. Furthermore, although a proactive detection of symptoms is supported to have positive association for road safety by recent studies (Beratis et al., 2017; Papadimitriou et al., 2017), machine learning techniques which have become popular due to their efficiency in a variety of domains, have also not been applied for cognitive disorder prediction. Therefore, this study will attempt to bridge the gap in the literature by applying machine learning classification algorithms for predicting if highly disaggregated vehicle motion characteristics belong to a depressed driving session and compare classification performance with time-series classifiers.

3. Methodology

As mentioned before, the purpose of this study is the prediction of sessions belonging to depressed drivers using microscopic driving simulator data. The underlying problem of distinguishing between healthy drivers and depressed individuals is a binary classification one, and therefore binary classification algorithms are going to be utilized.

3.1. Binary classification and performance evaluation metrics

In binary classification, the objective is to efficiently determine the "class" of unlabelled data instances, based on training examples from a labelled dataset, which is used as an example for the developed classifiers. Hence, in order to develop an efficient classifier, a training dataset $X_{training} = \{(x_n, y_n), n = 1, \dots N\}$ is considered, where x_n is a predictor variable and y_n=\{0,1\} is the response. In the current paper predictor variables are going to be described by microscopic motion characteristics obtained from a driving simulator, and the response variable is going to be described by the tuple {'Control', 'Depression}.

Classification performance of developed classifiers is initially assessed through the confusion matrix, which essentially compares and contrasts the original and predicted (i.e. after classification) label or class of a data instance, so as to verify their correct classification. Based on the confusion matrix, several performance metrics can be obtained, such as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \qquad (1)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \qquad (2)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad (3)$$

$$\text{G-means}= \sqrt{Recall * Specificity} \qquad (4)$$

$$\text{f1-measure}= \frac{2*precision*recall}{precision+recall} \qquad (5)$$

$$\text{False alarm rate}= \frac{\text{FP}}{\text{TN+FP}} \qquad (6)$$

where: TN: True Negative, TP: True Positive, FN: False Negative, FP: False Positive.

The terminology of "negative" and "positive" instances correlates to the distinction between the two classes. In the majority of problems, the class of interest (i.e. the class that is seek to be discovered among the data) is termed as the "positive" class, while a "negative" class describes the remaining data instances. In the current problem formulation, as the two classes are "Depression" and "Control", and "Depression" is the one needed to be efficiently identified, it forms the “positive” class. Consequently, data instances labelled as “Control” belong to the “negative” class.

As far as classification metrics are concerned, recall demonstrates the correct classification accuracy with respect to depressed driving sessions, while the specificity statistic shows the performance of the classifier in terms of the control cases. Precision is used for identifying the proportion of driving sessions predicted as belonging to “depressed” participants, and actually belong to “depressed”. G-means, essentially the geometric mean of recall and specificity, is used to ensure classification accuracy among both classes, even when there is an imbalance between their data instances. Lastly, the f1-measure is the harmonic mean of precision and recall and resembles the overall correct classification of “depressed” driving sessions (Tharwat, 2018).

3.2. Classification algorithm and the problem of imbalance

In order to be able to distinguish efficiently between depressed and healthy control driving sessions, an efficient algorithm needs to be utilized, with proven results in relevant domains. Random Forests (RFs) have been proven to perform well in studies that correlate driving simulator with driving behavior and safety assessment using both aggregated and time-series simulation measurements (Katrakazas et al., 2018; Katrakazas et al., 2019). Furthermore, they have been successfully applied in detecting depression among the medical society (Cacheda et al., 2019; Wade et al., 2015). Therefore, RFs show potential in order to be able to detect depression from driving simulator variables using disaggregated data and time-series measurements.

RFs are an ensemble classifier and more specifically a bagging algorithm. Bagging algorithms make use of only one learning algorithm and modify the training set by using the bagging algorithm to create new training sets (Breiman, 2001). RFs are an enhancement of bagged trees and utilize the bagging algorithm along with the random subspace method proposed by Ho (1998). Every tree is built using the impurity Gini index (Breiman, 2001), but only a random subset of the input features is used for constructing the tree without pruning. For the training dataset, one-third of the samples is randomly neglected and forms the so-called “out-of-bag” (OOB) samples, while the samples that are accepted are used for building the tree. For every constructed tree the OOB samples are used as a validation dataset and the misclassification OOB error is estimated. When a new data record needs to be assigned to a class, its attributes are run through the constructed trees and a classification result for every tree is obtained. The majority vote over all the classification results (i.e. from all trees) is chosen as the classified label for that specific data record (Verikas et al., 2011). However, an appropriate value for the number of features used for splitting a node of a tree needs to be tuned by the user in order for the OOB

One of the limitations in approaches concerned with identification of a (cognitive) impairment, and generally in disease identification is the documented imbalance of the utilized datasets. In disease identification, usually the data instances representing the class of healthy controls is overpopulated, while the class of patients has few examples. This is a well-documented problem in studies on machine learning and detection of depression or other MCIs (Bertoncello, M., Wee, 2015; Dipnall et al., 2016; Gerych et al., 2019; Munteanu et al., 2015). The inherent limitation of performing classifications on imbalanced datasets is the high misclassification rate for the patient class, as the algorithms favour the majority class, for which they have many example instances.

To overcome this problem, machine learning and data mining experts, propose data sampling, algorithm alteration or cost-sensitive learning (He and Garcia, 2009; López et al., 2013). The first and simplest solution (i.e. data sampling), describes the procedure of undersampling the majority class or synthetically oversampling the minority class in order to produce a “balanced” dataset, train a classifier on it, and test its results on the original and imbalanced dataset. The problem with oversampling according to (López et al., 2013) is that it usually leads to overfitting. However, if both approaches (i.e. under- and oversampling) are combined, they can provide an alternative solution to overcome individual limitations of the two techniques (Lemaitre et al., 2016).

Reviewing the literature in undersampling and overasampling with data cleansing, it was found that Edited Nearest Neighbours (ENN) (Wilson, 1972), and its integration with Synthetic Minority Oversampling TEchnique (SMOTE) (Chawla et al., 2002) performed well for classes that are difficult to recognise. This approach has also shown enhanced results, when applied in real-time collision prediction (Katrakazas et al., 2019; Katrakazas et al., 2019; Katrakazas, 2017), or safety assessment of driving simulator data. The procedure of SMOTE-ENN is as follows: After artificially generating instances of the minority class through SMOTE, ENN is implemented to conduct the data cleaning in depth and removes data instances from both classes when the three nearest neighbours of a data instance are misclassified. This is beneficial, especially for datasets with a small number of instances in the positive class (as in the depressed driving classification problems. As it can be observed, SMOTE-ENN forms another potential solution for the identification of depression among drivers using driving simulator data. The algorithm will be henceforth termed as SMOTE-ENN.

3.3. Aggregated data versus time-series modelling

As the objective of the study will be carried out using microscopic data from driving sessions, there needs to be a distinction between the levels of data aggregation. Among the transportation research domain, traffic data used for behaviour prediction or safety assessment are usually aggregated (Abdel-Aty et al., 2005; Franke and Krems, 2013), in order for post-trip or post-event interventions to be applied, while real-time applications (Habtemichael and Santos, 2012; Vlahogianni and Barmpounakis, 2017) demand the use of highly disaggregated or time-series data, in order to identify different behaviours or critical events in a very short time horizon. In order for the present study, to be consistent with the state-of-the-art in both approaches, data from the driving simulator will be aggregated in short time intervals and simultaneously the time-series of the recorded variables will be explored so as to investigate the most effective approach. The data aggregation intervals as well as the time-series will have a length of 30-seconds, 1-minute and 5-minutes, as these intervals and time-series lengths have been also utilized in previous case-control traffic safety studies (Katrakazas et al., 2019; Katrakazas et al., 2017).

4. Data Description and Pre-processing

The data utilized in this study were collected using a driving simulator at the Department of Transportation Planning and Engineering of the National Technical University of Athens. More specifically, a FOERST Driving Simulator FP, consisting of 3 LCD wide screens 40'' (Full HD: 1920x1080 pixels, a driving position and a support motion base was employed. The simulator's dimensions at full development are 230x180cm, the width of its base is 78cm and its total field of view is 170 degrees. The data collected with the simulator were originally used for the Distract and DriverBrain projects (Pavlou, 2016; Yannis et al., 2014) which investigated the causes and impacts of driver distraction, as well as the driving capabilities of drivers with MCIs using a driving simulator.

With regards to the MCI drivers, all of the participants were recruited among the patients of the 2nd Department of Neurology of the University of Athens Medical School at ATTIKON University General Hospital, after informed consent was obtained from all individuals. Participants, were initially evaluated on a full clinical medical, ophthalmological and neurological scope in order to document the characteristics of their disorders, as well as complementary parameters with a potential impact on driving (e.g. medication). A second assessment also took place with regards to neuropsychological tests and psychological behavioural questionnaires. With regards to depression, patients were evaluated using the Patient Health Questionnaire (PHQ-9) scale (Cameron et al., 2008). According to the PHQ-9 scale, participants were found to have minimal, mild and moderate depression (Beratis et al., 2017), which corresponds to scores 1-14 on the scale. More specifically, 65% of the control group had none or minimal depression (PHQ-9 values 0-4) and 35% had mild symptoms (PHQ-9 values 5-9), while in the depressed group 18.18% had minimal (PHQ-9 values 0-4), 54.55% mild (PHQ-9 values 5-9) and 27.27% moderate depression (PHQ-9 values 10-14).

The driving scenarios included driving in rural, urban and motorway environments. For the analysis part of this paper only the rural area data were used, as these were available to the authors. Each experiment included a 15- to 20-minute warm-up drive, so as to familiarize the driver with the simulator, and a 20-minute recorded driving session. The rural route was 2.1 km long on a single carriageway, with 3m lane width, zero gradient and mild horizontal curves. During each trial, 2 unexpected incidents were programmed to occur and concerned the sudden appearance of an animal. The experiment was counterbalanced with regards to the number and order of trials. For more details on the dataset, the experiment and the medical evaluation of participants, the reader is referred to (Beratis et al., 2017; Pavlou, 2016; Yannis et al., 2014).

Measurements from the driving simulator, were recorded every 17 and 33 milliseconds, and the variables of interest, in order to predict depression, were chosen to be speed, time to headway (i.e. collision with the ahead driving vehicle), TTC, as well as lateral and longitudinal acceleration. In order for the classifiers to be developed, driving sessions of healthy controls and depressed drivers were chosen. In total, driving sessions of 11 depressed drivers and 65 healthy controls were chosen, adding up to 2,700,223 raw measurements. The participants included in the final dataset, were not under treatment with antidepressant medication. It should also be mentioned here that the final dataset for classification contained labelled data from all drivers (both depressed and control), in order for the models to be able to distinguish between sessions belonging to depressed or control participants. The details on the age, education and driving experience of the drivers are summarized in Table 1. Depressed and the control samples were matched with regards to age, so as to get more clear classification results. Matching was performed by taking control samples which are in the range of two standard deviations around the mean of the age of the depressed sample. Due to the large number of observations, the raw observations were aggregated initially to 250millisecond intervals, so as to reduce noise in the dataset. After obtaining the aggregated 250 ms measurements and in order to obtain the mean and standard deviations for 30-seconds, 1-minute and 5-minute intervals, consecutive time "segments" of 30-second, 1-minute and 5-minute duration were used to filter measurements and acquire the corresponding statistics for each duration. In order to avoid the development of classifiers with correlated variables, the correlation heatmap was estimated for each aggregation interval (i.e. 30-seconds, 1-minute and 5-minute data) across all data points and is presented in Figure 1.

Table 1: Description of the control and depressed drivers

Variable	Depressed (N=11)				Control(N=65)			
	mean	std	min	max	mean	std	min	max
Age (years)	55.61	11.72	27.00	73.00	54.66	11.47	33.00	78.00
Education (years)	11.04	3.19	6.00	15.00	15.30	3.16	6.00	24.00
Driving_experience (years)	31.69	12.86	5.00	55.00	31.26	10.60	9.00	50.00

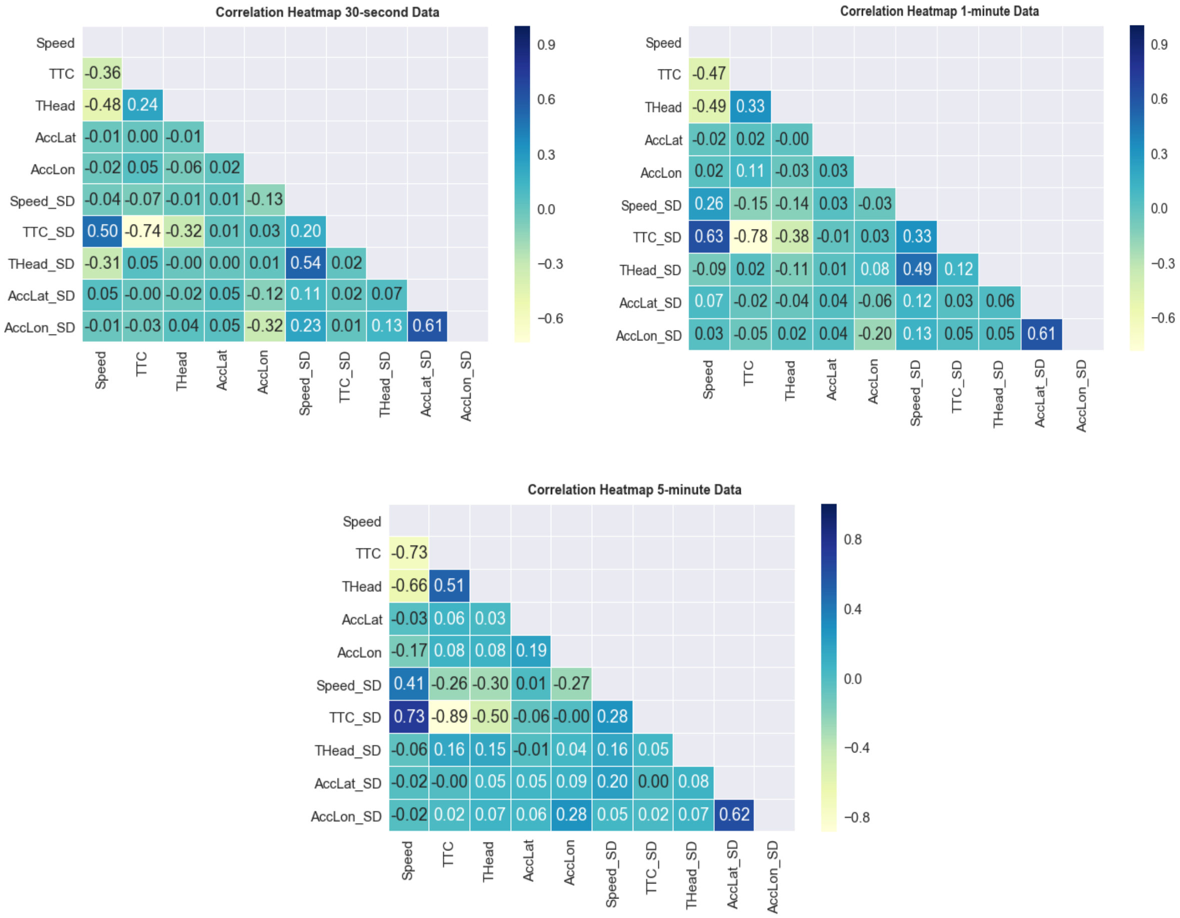

Figure 1: Correlation heatmaps of independent variables from the driving simulator (top left: 30-seconds data, top right: 1-minute data and bottom: 5-minute data)

From Figure 1, it can be observed that surrogate safety indicators (i.e. TTC and Time to Headway) are highly correlated with each other and with speed, and as a result it was decided to drop these variables and include only the kinematic characteristics of driving for the remaining of the analysis, i.e. the mean and standard deviation of speed, lateral acceleration and longitudinal acceleration and as a result six variables were utilized for the development of the classifiers on aggregated data, and the time-series analysis. Descriptive statistics for the three general variables (i.e. speed, lateral acceleration and longitudinal acceleration) are given in Table 2.

Table 2: Descriptive statistics for the included variables

Variable	Depressed (N=11)				Control(N=65)			
	mean	std	min	max	mean	std	min	max
Speed (km/h)	30.92	19.59	0.00	119.50	34.85	20.55	0.00	104.80
AccLat (m/s²)	-0.04	20.00	-11243.00	32.08	0.01	13.69	-6826.00	8007.00
AccLon (m/s²)	27.28	1150.54	-535182.00	6919.00	31.62	1430.082	-528154.00	7886.00

In order for the variable length to be extracted, the 250-millisecond observations were split in the corresponding intervals (i.e. 30 seconds, 1 minute and 5 minutes) and each series of observations was labelled according to the driver's psychological status (i.e. 'control' or 'depression). As a result, six time-series (one for each of the mean and standard deviation of speed, lateral and longitudinal acceleration) were created for each time interval, adding up to 21 datasets in total (6 time-series * 3 intervals + 3 datasets with aggregated observations for each interval). For each time-series dataset, and in order to optimize the classification results, rows (i.e. individual time-series) were deleted if the majority of their observations were zero. Furthermore, the total number of cases for control and depressed driver observations was extracted, so as to identify if an imbalanced learning technique (e.g. SMOTE-ENN) should be applied. The extracted ratio of control:depressed data instances in each of the datasets was:

- Aggregated data in 30-seconds intervals 1:7
- Aggregated data in 1-minute intervals 1:7
- Aggregated data in 5-minute intervals 1:6
- 30-seconds time-series 1:6
- 1-minute time-series 1:6.
- 5-minute time-series 1:6

Consequently, as there is a difference between the instances of the two classes, it was decided that SMOTE-ENN will be utilized along RFs in order to obtain balanced classification performance

5. Results and Discussion

As discussed in the methodology section, both the aggregated datasets, as well as the time-series were attempted to be classified using the RF algorithm, which is a powerful ensemble classifier, and the assistance of SMOTE-ENN to investigate any potential enhancement from an imbalanced learning technique. Before the initiation of each algorithm, an optimization routine was run along with 10-fold cross-validation in order to find the optimal parameters for RF classifiers. More specifically, for each RF, the number of estimators, the maximum depth of the tree, and the maximum number of features to consider when looking for the best split of a node were optimized. In order to avoid over-fitting and assure the validity of the results, 70% of the dataset was used for training the classifiers and the remaining 30% was used for testing the classification results. The models were developed in Python 3.7 using the scikit-learn (Pedregosa et al., 2012) package for RFs and the *imbalanced-learn* package (Lemaitre et al., 2016) for the application of SMOTE-ENN with regards to imbalanced learning. SMOTE-ENN was run with SMOTE taking into account 10 data instances neighbours for generating synthetic samples and ENN cleaning the dataset using the three nearest neighbours of data cases. Each RF run with SMOTE-ENN for balancing the dataset, was trained on the balanced dataset (acquired after undersampling the majority class and oversampling the minority class) and the performance was tested on the original (imbalanced) dataset. By testing the performance on the original dataset, it is ensured that the validation of the classification results is not based on artificially created instances from SMOTE or a smaller sample acquired through ENN, but is directly acquired from the attributes of the original dataset.

The classification results, were evaluated through equations 1-6 for their scores in precision, recall, specificity, g-means, f1-score and false alarm rates. Figure 2 demonstrates the results for recall, specificity and false alarm rate for the developed classifiers. Recall, Specificity and False Alarms were chosen because it is essential for the classifiers to be able to predict correctly cases of depressed drivers (which is shown by the recall metric), cases of healthy control (given by the specificity measure) and not make misclassifications with regards to the mental status of drivers (as shown by the false alarm rate). Classifiers were accepted if they succeeded in sufficiently identifying both control and depressed driver data points with a false alarm rate lower than 30%. In Figure 2, every classification algorithm is marked according to the following coding: "RF_30" denotes the RF classifier on 30-seconds aggregated data, "RF_SMOTE-ENN_1" denotes the RF classifier treated with the imbalanced technique of SMOTE-ENN on 1-minute aggregated data, and "Speed_StdDev_SMOTE-ENN_1" denotes the RF classifier on the 1-minute time-series of the standard deviation of speed, treated with the SMOTE-ENN imbalanced learning technique. It should be noted here, that the classifiers for the 5-minute time-series of the standard deviation of speed using SMOTE-ENN, and the same classifiers for 5-minute time-series of mean and standard deviation of lateral and longitudinal acceleration did not converge and therefore their results are not presented. Furthermore, it was found that all time-series classifiers without the treatment for the imbalance of the dataset failed to correctly identify any of the depressed drivers, and therefore only results of the SMOTE-ENN treated RFs are presented in Figure 2. The same result (i.e. the failure to detect data points belonging to depressed drivers) was observed for the classifiers regarding aggregated data without treatment for imbalanced learning. This is probably due to the fact that the data imbalance between control and depressed driver cases is such that algorithms fail to identify depressed cases without any assistance from imbalanced learning.

From Figure 2, it can be observed that an efficient prediction of sessions belonging to drivers suffering from depression is indeed feasible. The accepted classifiers are able to predict correctly cases of depressed drivers, even with highly disaggregated driver behaviour data (i.e. data aggregated in 30-second intervals). Furthermore, it is demonstrated that the developed classifiers can make correct predictions regarding both classes, as both recall and specificity statistics are high among the models. Another important finding is that the developed classifiers, have succeeded in distinguishing between depressed and control driving sessions, with a very small percentage of false alarms. More specifically, the highest false alarm rate was found in data aggregated in 1-minute intervals (i.e. 20%), as well as the 30-second time-series of the mean lateral acceleration (i.e. 25%). These two models also have the lowest rates of specificity, which shows their limitation in correctly identifying health control cases. With regards, to treating aggregated data with imbalanced learning it is obvious that small differences are observed, with the best performance obtained using driving data aggregated in 5-minute intervals. With regards to individual time-series of driving indicators, the best results are obtained using 30-seconds and 1-minute measurements of mean speed, while the best results for 5-minute time series were obtained using the series of the standard deviation of the longitudinal acceleration but with 33% of false alarms. These indicators could be efficiently utilized in real-time applications in order to detect driving sessions of depressed drivers. When time-series of shorter duration are utilized, the best predictors for depressed driving sessions are the mean speed and the standard deviation of longitudinal accelerations. To further validate the results, Table 3 gives the full classification results for the ten best classifiers.

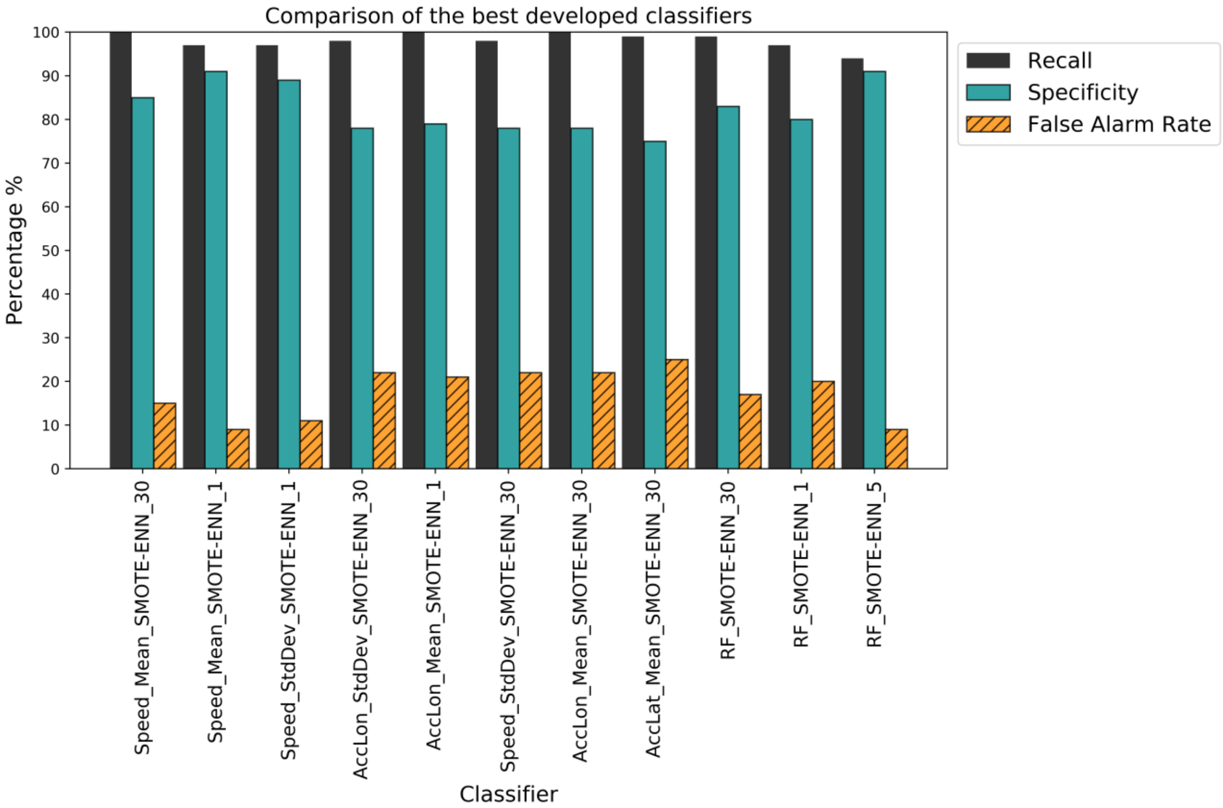

Figure 2: Classification performance for the best ten (a) and the rejected classifiers (b)

Table 3: Overall performance of the ten best developed classifiers

Classifier	Precision	Recall	Specificity	f1-score	G-Means	FP Rate
Speed_Mean_SMOTE-ENN_30	92.89%	99.81%	85.04%	96.23%	96.29%	14.96%
Speed_Mean_SMOTE-ENN_1	95.45%	96.92%	91.49%	96.18%	96.19%	8.51%
Speed_StdDev_SMOTE-ENN_1	95.24%	97.01%	89.26%	96.12%	96.12%	10.74%
AccLon_StdDev_SMOTE-ENN_30	93.67%	98.48%	77.56%	96.01%	96.04%	22.44%
AccLon_Mean_SMOTE-ENN_1	92.59%	99.60%	78.95%	95.97%	96.03%	21.05%
Speed_StdDev_SMOTE-ENN_30	92.16%	98.33%	77.72%	95.14%	95.19%	22.28%
AccLon_Mean_SMOTE-ENN_30	90.51%	99.61%	78.05%	94.84%	94.95%	21.95%
AccLat_Mean_SMOTE-ENN_30	91.11%	98.87%	74.75%	94.83%	94.91%	25.25%
RF_SMOTE-ENN_30	88.97%	99.18%	82.89%	93.80%	93.94%	17.11%
RF_SMOTE-ENN_1	88.12%	97.27%	80.49%	92.47%	92.58%	19.51%
RF_SMOTE-ENN_5	88.89%	94.12%	90.91%	91.43%	91.47%	9.09%

Table 3 further demonstrates the power of RFs treated with SMOTE-ENN in identifying depressed drivers in short-term time-series and aggregated data. Monitoring mean speed in 30-seconds or 1-minute time durations led to identifying almost all the depressed drivers among the healthy controls, with only up to 15% false alarms. Furthermore, the high scores of such classifiers in precision, f1 and G-means proves that the predictions are at their majority relevant and balanced among the two classes. Hence, both depressed and healthy drivers can effectively be predicted. As a result, from the developed models, it is found that mean longitudinal acceleration, speed and lateral accelerations are the three best indicators of depressed driving data points. As Table 3 shows that the best classification results are obtained using individual driving behaviour indicators such as speed and acceleration, it can be concluded that time-series outperform aggregated data in identifying depressed driving. This is probably a consequence of the fact that time-series data capture better the dynamic nature of the driving task, and includes useful information that is lost in the aggregation procedure.

In order to compare the results of this paper with similar studies in the field, the results of Papadimitriou et al., (2017) were used, as this study had a similar objective with our present study (i.e. the identification of cognitive impairment through driving simulator measurements). In Papadimitriou et al., (2017), which aimed at distinguishing between patients of MCI, Alzheimer's disease and healthy drivers, only up to 63% of MCI patients were identified, and up to 47% of Alzheimer's disease cases, percentages which are at best 30% lower than the majority of the classifiers developed in this paper. These findings validate the enhancement offered by machine learning approaches and the utilization of highly disaggregated time-series of driving simulator observations.

6. Conclusions

Depression is one of the most frequent mental disorders and has been found to be negatively correlated with driving performance, usually among elderly drivers. Although there is a plethora of studies investigating the effects of depression on driving behavior indices, there is a significant gap in studies seeking to identify indicators of depression from driving attributes and especially using machine learning techniques. The present study, takes a first step in bridging that gap, by developing Random Forests classifiers, treated with imbalanced learning and trained on highly disaggregated (30-seconds, 1-minute and 5-minute) driving behavior measurements and corresponding time-series.

Results of the developed models, are more than promising, and demonstrate that time-series of mean longitudinal acceleration and speed can be utilized to identify drivers with depression, with a very low percentage of false alarms, even with 30-seconds observations. In a current transportation environment, that is rapidly becoming automated, this finding is extremely important, as AVs could "sense" a potential psychological disorder through sequential vehicle kinematic observations and take control of the vehicle to assure more comfort and safety for the passengers.

Nevertheless, the current study is just a preliminary investigation on how machine learning can be utilized in predicting mental disorders from driver analytics. Deep learning, naturalistic driving data and multiple classification of other cognitive and mental disorders are envisioned to enhance the findings of the current paper. Moreover, in order for a classification approach similar to the one developed in the current paper to become more person-based, the classified aggregated and time series driving data should be contrasted with the remaining data of the same driver in order to more clearly identify if a specific driver is depressed.

Future research should focus on exploiting similar classification results for the development of human-machine interfaces (HMIs) that could assist depressed drivers or drivers with other cognitive impairments while driving. On the same principle, personalized in-vehicle or post-trip safety interventions (e.g. coaching or gamification) tailored to the needs of drivers suffering from depression could be developed. Moreover, driving simulator experiments exclusively on depressed patients and the assistance of psychologists and psychiatrists in developing depression-targeted scenarios could become beneficial for developing even more powerful classifiers to detect depression from driving sessions.

As a final note, it should be noted that although prediction of depression is sufficiently supported from the findings of the paper, in no case, should driving analytics circumvent neuropsychological assessments. However, they could act as a precursor or a complementary form of assessing mild psychological or mental disorders.

References

Abdel-Aty, M., Pande, A., Hsia, L.Y., Abdalla, F., 2005. The Potential for Real-Time Traffic Crash Prediction. In: ITE Journal on the Web. pp. 69–75.

Ballenger, J.C., 2008. Depression in Patients With Mild Cognitive Impairment Increases the Risk of Developing Dementia of Alzheimer Type: A Prospective Cohort Study. Yearb. Psychiatry Appl. Ment. Heal. 2006, 268–269.

Beratis, I.N., Andronas, N., Kontaxopoulou, D., Fragkiadaki, S., Pavlou, D., Papatriantafyllou, J., Economou, A., Yannis, G., Papageorgiou, S.G., 2017. Driving in mild cognitive impairment: The role of depressive symptoms. Traffic Inj. Prev. 18, 470–476.

Bertoncello, M., Wee, D., 2015. Ten Ways Autonomous Driving Could Redefine the Automotive World. Market Evaluation on Development of Autonomous Vehicles.

Breiman, L., 2001. Random Forests. Mach. Learn. 45.1, 5–32.

Brunnauer, A., Laux, G., 2017. Driving under the Influence of Antidepressants: A Systematic Review and Update of the Evidence of Experimental and Controlled Clinical Studies. Pharmacopsychiatry 50, 173–181.

Bulmash, E.L., Moller, H.J., Kayumov, L., Shen, J., Wang, X., Shapiro, C.M., 2006. Psychomotor disturbance in depression: Assessment using a driving simulator paradigm. J. Affect. Disord. 93, 213–218.

Cacheda, F., Fernandez, D., Novoa, F.J., Carneiro, V., 2019. Early Detection of Depression: Social Network Analysis and Random Forest Techniques. J. Med. Internet Res. 21, e12554.

Cameron, I.M., Crawford, J.R., Lawton, K., Reid, I.C., 2008. Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care. Br. J. Gen. Pract. 58, 32–36.

Chawla, N. V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.

Cunningham, M.L., Regan, M.A., 2016. The impact of emotion, life stress and mental health issues on driving performance and safety. Road Transp. Res. 25, 40–50.

Dipnall, J.F., Pasco, J.A., Berk, M., Williams, L.J., Dodd, S., Jacka, F.N., Meyer, D., 2016. Into the bowels of depression: Unravelling medical symptoms associated with depression by applying machine-learning techniques to a community based population sample. PLoS One 11, 1–19.

Fiske, A., Loebach Wetherell, J., Gatz, M., 2012. Depression in older adults. Am. J. Nurs. 112, 22–30.

Franke, T., Krems, J.F., 2013. Understanding charging behaviour of electric vehicle users. Transp. Res. Part F Traffic Psychol. Behav. 21, 75–89.

Gerych, W., Agu, E., Rundensteiner, E., 2019. Classifying Depression in Imbalanced Datasets Using an Autoencoder- Based Anomaly Detection Approach. Proc. - 13th IEEE Int. Conf. Semant. Comput. ICSC 2019 124–127.

Grahek, I., Shenhav, A., Musslick, S., Krebs, R.M., Koster, E.H.W., 2019. Motivation and cognitive control in depression. Neurosci. Biobehav. Rev. 102, 371–381.

Habtemichael, F.G., Santos, L. de P., 2012. The Need for Transition from Macroscopic to Microscopic Traffic Management Schemes to Improve Safety and Mobility. Procedia - Soc. Behav. Sci. 48, 3018–3029.

He, H., Garcia, E.A., 2009. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21, 1263–1284.

Hill, L.L., Lauzon, V.L., Winbrock, E.L., Li, G., Chihuri, S., Lee, K.C., 2017. Depression, antidepressants and driving safety. Inj. Epidemiol. 4.

Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20, 832–844.

Katrakazas, C., 2017. Developing an advanced collision risk model for autonomous vehicles 272.

Katrakazas, C, Antoniou, C., Yannis, G., 2019. Time Series Classification Using Imbalanced Learning for Real-Time Safety Assessment 1–15.

Katrakazas, C., Quddus, M., Chen, W.H., 2017. A Simulation Study of Predicting Real-Time Conflict-Prone Traffic Conditions. IEEE Trans. Intell. Transp. Syst. 1–12.

Katrakazas, C., Quddus, M., Chen, W.H., 2018. A simulation study of predicting real-time conflict-prone traffic conditions. IEEE Trans. Intell. Transp. Syst. 19, 3196–3207.

Katrakazas, Christos, Quddus, M., Chen, W.H., 2019. A new integrated collision risk assessment methodology for autonomous vehicles. Accid. Anal. Prev. 127, 61–79.

Lemaitre, G., Nogueira, F., Aridas, C.K., 2016. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. CoRR abs/1609.0, 1–5.

Levinson, J., Askeland, J., Becker, J., Dolson, J., Held, D., Kammel, S., Kolter, J.Z., Langer, D., Pink, O., Pratt, V., Sokolsky, M., Stanek, G., Stavens, D., Teichman, A., Werling, M., Thrun, S., Hardware, A., 2011. Towards Fully Autonomous Driving : Systems and Algorithms.

López, V., Fernández, A., García, S., Palade, V., Herrera, F., 2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Inf. Sci. (Ny). 250, 113–141.

McDonald, C.C., Sommers, M.S., Fargo, J.D., Seacrist, T., Power, T., 2018. Simulated Driving Performance, Self-Reported Driving Behaviors, and Mental Health Symptoms in Adolescent Novice Drivers. Nurs. Res. 67, 202–211.

Mojtabai, R., Olfson, M., Han, B., 2016. National Trends in the Prevalence and Treatment of Depression in Adolescents and Young Adults. Pediatrics 138, e20161878–e20161878.

Munteanu, C.R., Fernandez-Lozano, C., Mato Abad, V., Pita Fernández, S., Álvarez-Linera, J., Hernández-Tamames, J.A., Pazos, A., 2015. Classification of mild cognitive impairment and Alzheimer’s Disease with machine-learning techniques using 1 H Magnetic Resonance Spectroscopy data. Expert Syst. Appl. 42, 6205–6214.

Papadimitriou, E., Yannis, G., Pavlou, D., Beratis, I., Papageorgiou, S.G., Transportation Research, B., 2017. Can Driving in the Simulator Diagnose Cognitive Impairments? Transp. Res. Board, 96th Annu. Meet. 14p.

Pavlou, D., 2016. Traffic and safety behaviour of drivers with neurological diseases affecting cognitive functions.

Pavlou, D., Beratis, I., Papadimitriou, E., Antoniou, C., Yannis, G., Papageorgiou, S., 2016. Which Are the Critical Measures to Assess the Driving Performance of Drivers with Brain Pathologies? Transp. Res. Procedia 14, 4393–4402.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2012. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Scott-Parker, B., Watson, B., King, M.J., Hyde, M.K., 2013a. A further exploration of sensation seeking propensity, reward sensitivity, depression, anxiety, and the risky behaviour of young novice drivers in a structural equation model. Accid. Anal. Prev. 50, 465–471.

Scott-Parker, B., Watson, B., King, M.J., Hyde, M.K., 2013b. A further exploration of sensation seeking propensity, reward sensitivity, depression, anxiety, and the risky behaviour of young novice drivers in a structural equation model. Accid. Anal. Prev. 50, 465–471.

Tharwat, A., 2018. Classification assessment methods. Appl. Comput. Informatics.

Vardaki, S., Yannis, G., Antoniou, C., Pavlou, D., Beratis, I., Papageorgiou, S.G., 2014. DO SIMULATOR MEASURES IMPROVE IDENTIFICATION OF OLDER DRIVERS WITH MCI? In: 94th Annual Meeting of the Transportation Research Board, Washington, January 2015.

Verikas, A., Gelzinis, A., Bacauskiene, M., 2011. Mining data with random forests: A survey and results of new tests. Pattern Recognit. 44, 330–349.

Vlahogianni, E.I., Barmpounakis, E.N., 2017. Driving analytics using smartphones: Algorithms, comparisons and challenges. Transp. Res. Part C Emerg. Technol. 79, 196–206.

Wade, B.S.C., Joshi, S.H., Pirnia, T., Leaver, A.M., Woods, R.P., Thompson, P.M., Espinoza, R., Narr, K.L., 2015. Random forest classification of depression status based on subcortical brain morphometry following electroconvulsive therapy. Proc. - Int. Symp. Biomed. Imaging 2015-July, 92–96.

Wickens, C.M., Smart, R.G., Mann, R.E., 2014. The Impact of Depression on Driver Performance. Int. J. Ment. Health Addict. 12, 524–537.

Wilson, D.L., 1972. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. IEEE Trans. Syst. Man Cybern. 2, 408–421.

World Health Organization, 2017. Depression and Other Common Mental Disorders. Institutes Heal. Natl. 1–22.

Yannis, G., Golias, J., Antoniou, C., Vardaki, S., Papantoniou, P., Pavlou, D., Espié, S., Kalisperakis, G., Papageorgiou, S.G., Tsivgoulis, G., Bonakis, A., Andronas, N., Papatriantafyllou, I., Liozidou, A., Kontaxopoulou, D., Economou, A., Kosmidis, M., 2014. Distract: Causes and Impacts of Driver Distraction: A Driving Simulation Study, Deliverable 4: Driving Simulator Experiment.