Chelsea Dobbins and Stephen Fairclough, "Detecting Negative Emotions During Real-Life Driving via Dynamically Labelled Physiological Data" in 2018 IEEE International Conference on Pervasive Computing and Communications (PerCom'18), Athens, Greece, 19th – 23rd March, 2018 (Accepted)

Detecting Negative Emotions During Real-Life Driving via Dynamically Labelled Physiological Data

Chelsea Dobbins
Department of Computer Science
Liverpool John Moores University
Liverpool, United Kingdom
C.M.Dobbins@ljmu.ac.uk

Stephen Fairclough
School of Natural Sciences and Psychology
Liverpool John Moores University
Liverpool, United Kingdom
S.Fairclough@ljmu.ac.uk

***Abstract*— Driving is an activity that can induce significant levels of negative emotion, such as stress and anger. These negative emotions occur naturally in everyday life, but frequent episodes can be detrimental to cardiovascular health in the long term. The development of monitoring systems to detect negative emotions often rely on labels derived from subjective self-report. However, this approach is burdensome, intrusive, low fidelity (i.e. scales are administered infrequently) and places huge reliance on the veracity of subjective self-report. This paper explores an alternative approach that provides greater fidelity by using psychophysiological data (e.g. heart rate) to dynamically label data derived from the driving task (e.g. speed, road type). A number of different techniques for generating labels for machine learning were compared: 1) deriving labels from subjective self-report and 2) labelling data via psychophysiological activity (e.g. heart rate (HR), pulse transit time (PTT), etc.) to create dynamic labels of high vs. low anxiety for each participant. The classification accuracy associated with both labelling techniques was evaluated using Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM). Results indicated that classification of driving data using subjective labelled data (1) achieved a maximum AUC of 73%, whilst the labels derived from psychophysiological data (2) achieved equivalent performance of 74%. Whilst classification performance was similar, labelling driving data via psychophysiology offers a number of advantages over self-reports, e.g. implicit, dynamic, objective, high fidelity.**

Keywords—Mobile/Wearable Devices; Pervasive Computing; Emotion Recognition; Classification; Driving; Labelling

I. INTRODUCTION

Driving is a common activity that millions of people undertake daily. However, the accumulation of time spent in motor vehicles over years and decades is significant – for example, commuters in the United Kingdom spend an average of 52 minutes per day commuting to and from work [1], which equates to 4.3 hours per week, 17.3 hours per month or 208 hours per year, whilst those in the United States report spending an average of 47.1 minutes driving daily [2]. Driving is also a significant source of everyday stress that adversely affects health by increasing arousal, heart rate, and blood pressure [3]. However, most people undertake this seemingly innocuous activity with little thought about the long-term impact of driving on health.

Stress and anger are negative emotions that naturally occur in everyday life. However, excessive, persistent and repeated exposure to stress can result in lasting and harmful effects on the physical and mental health of the individual [4]. For example, excessive stress can lead to headaches, insomnia and fatigue, whilst long term exposure over years and decades is associated with increased risk of chronic diseases, including cardiovascular disease (CVD) [5]. CVD is a group of disorders that affects the heart and blood vessels, including coronary heart disease that can lead to acute events, such as heart attacks and strokes [6]. CVD is the leading cause of death globally, with 17.7 million deaths in 2015 [6]. Nevertheless, stress can be managed by developing effective coping strategies, which can promote growth, adaptation and resilience to the deleterious impact of stress on health [7].

The measurement of negative emotions outside of the laboratory places significant reliance on self-reported levels of stress [5]. This approach is limited because self-report methods are: subjective to bias, intrusive and are only sampled at a low rate, such as 3-4 times a day [5]. In contrast, mobile and wearable technology provide a continuous stream of quantitative data of psychophysiological stress (e.g. elevated heart rate, increased skin conductance level, etc.) without requiring any action from the participant [8].

Machine learning algorithms can be used to detect and classify negative emotions in everyday life. However, these algorithms are dependent on labelling data in a valid fashion, i.e. accurately representing a distinction between two or more psychological states. In most cases, these labels are derived on the basis of subject self-report data [9]. Whilst self reports are useful for capturing subjective aspects of emotion, they can be problematic for data collection outside of the laboratory, as they impose significant burden on the participant [5], [9].

This work will investigate the issue of labelling data sets by comparing classification accuracy when labels derived from subjective and psychophysiological data are applied to driving data. The aim of exploring labels derived from psychophysiology was to provide a viable option for data labelling *without* using subjective questionnaires (standard approach). This is an important issue because key features of the driving environment or transitory events (high traffic density, journey impedance, infrastructure, etc.) are known triggers for driver stress [10], [11]. Therefore, we can use subjective labels to differentiate a stressful drive from a non-stressful drive as perceived overall by the person. However, certain stressful events may occur momentary, such as traffic jams, another vehicle unexpectedly overtaking etc.). In these cases, subjective labels do not have sufficient resolution to capture these momentary events. Hence, we have explored the use of labels

This work has been supported by the UK Engineering and Physical Sciences Research Council (EPSRC) under Research Grant EP/M029484/1.

derived from psychophysiology, which are captured over the entirety of the journey and at a higher level of resolution.

The remainder of this paper is constructed as follows. Section 2 describes related work within the field of mobile emotion sensing. Section 3 presents two case studies, as well as our labelling approach. Section 4 illustrates results that have been obtained from comparing the subjective and physiological labels in classifying driving data. Section 5 provides a discussion of the results before the paper is concluded in section 6.

II. Related Work

Previous work used a variety of data sources, signal processing and machine learning approaches for stress detection [12]–[15] For instance, Hovsepian et al. [12] have utilized electrocardiogram (ECG), respiration data and derived labels from self-report during laboratory and field studies to detect stress using a Support Vector Machine (SVM) algorithm. Their model achieved a median accuracy of 72%. In other work, Muaremi et al. [13] detected stress using sleeping patterns. Their work captured self-reports and data from ECG, heart rate variability (HRV), respiration, body temperature, galvanic skin response (GSR) and accelerometer data to classify the results using SVM, logistic regression (Logit), k-nearest-neighbour (kNN), random forest (RF), and neuronal network (NN). Stress scores from the self-reports were used to obtain labels for the classification, which achieved a 73% accuracy in distinguishing between low, moderate, and high stress using SVM.

Alternatively, Garcia-Ceja et al. [14] measured stress in working environments using smartphone accelerometers and self-reported questionnaires. Using a combination of statistical models to classify self-reported stress levels achieved an overall accuracy of 71%. Similarly, Muaremi et al. [15] aimed to measure stress throughout the workday and during sleep. Their approach deployed a smartphone to collect data during the day, including self-reports, accelerometer, GPS, microphone, etc., whilst a Wahoo chest belt collected HRV data during sleep. Their classification analyses achieved 55% accuracy using only the smartphone data, 59% using only HRV features and 61% when all the features were combined. The stress scores from the self-reports were used to obtain labels for the classification.

III. Case Studies

Two studies were undertaken to collect real-life data from participants during their daily commuter journeys to and from their place of work.

A. Participants

Study 1 included thirteen participants – seven females and six males (mean = 41.69 years, SD = 11.75). Study 2 included eight participants – six females and two males (mean = 39.50 years, SD = 11.10). None of the participants had any history of heart disease and none were taking medication that would influence cardiovascular activity. The University Ethical Committee approved all procedures for participant recruitment and data collection prior to commencement of data collection.

Data were collected using our mobile data collection platform, which consisted of a smartphone and three Shimmer3™ sensors, including a 5-lead electrocardiography (ECG) unit, an optical pulse ear photoplethysmogram (PPG) clip and accelerometer. The ear was chosen because it provided a relatively stable site for data collection, as opposed to other areas, such as the fingertip, which is highly susceptible to motion artefacts [16]. Before commencement of the studies, participants were briefed with a description of the task and were trained with the equipment.

B. The Driving Tasks

The data collection protocol involved gathering data from participants over five working days, during their normal driving commutes to and from work. It was required that the commuter journey was: at least 10 minutes in duration, utilized the same route to/from work at approximately the same time, that the driver was alone in the car and did not listen to music during the journeys.

Wearable Shimmer3™ sensors were utilized across both studies to capture raw electrocardiography (ECG) and photoplethysmogram (PPG) signals. During the first study, a Shimmer3™ accelerometer was affixed in a flat position inside the vehicle to capture acceleration, which was later converted into features of speed. The Shimmer3™ sensors were configured to a sample rate of 512 Hz and data were stored on the internal microSD card of each device.

During study 2, a greater variety of driving data were obtained, which included location and road infrastructure (number of lanes, type of road, traffic density, etc.). A custom-built Android application, running on a Samsung™ Galaxy S5/S6 smartphone, was developed to capture first-person photographs of the environment every 30 seconds. The photographs were used to illustrate the road infrastructure (e.g. number of lanes, road type, etc.), as well as the dynamics of the drive (e.g. traffic density). Photographs were captured every 30 seconds by placing the phone into a mobile phone holder so that photographs could be taken of the driver view, via the windshield. Speed and location (latitude/longitude) data were also captured via this application.

Subjective reporting of stress was also obtained using a short-version of the State–Trait Anger Expression Inventory 2 (STAXI 2) [17] (study 1). Responses were measured in terms of state anger (S-Ang), feeling angry (S-Ang/ F), feeling like expressing anger verbally (S-Ang/V) and feeling like expressing anger physically (S-Ang/P). The state anger (S-Ang) scale was used during this work as this refers to the intensity of the individual's angry feelings [18]. However, due to issues of social desirability (i.e. participants were reluctant to report anger during their journeys), the STAXI 2 questionnaire was replaced in study 2 with a short-version of the UWIST Mood Adjective Checklist (UMACL) [19], which was used to capture subjective changes in mood due to each journey. This questionnaire has three major dimensions: energetical arousal (alert vs. tired), tense arousal (anxious vs. relaxed) and hedonic tone (happy vs. sad). The tense arousal scale was used during this work. Participants were required to complete both questionnaires *before* and *after* each journey to account for any changes in anger/mood that occurred during the duration of the drive. The questionnaires were administered using a custom-built Android application on a Samsung™ Galaxy S5/S6 smartphone.

The average driving time during study 1 was 32:12 min and study 2 was 37:22 min. The minimum driving time during study 1 was 10:44 min, whilst study 2 was 16:30 min. The maximum driving time was 77:45 min during study 1 and 108:30 min during study 2. Study 1 resulted in the collection of 366,200,928 samples of raw data (62:47:53 hours). Study 2 resulted in the collection of 159,496,783 instances of raw data (42:57:47 hours). In total, 525,697,711 (105:45:40 hours) instances of raw data have been collected across both data collection exercises.

C. Data Pre-Processing

Collection of psychophysiological data in the field is particularly susceptible to noise and data loss [12]. Data loss and distortion can occur for a multitude of reasons, including loss of contact with the sensors and the influence of physical movement. Therefore, it is important that sensor data is pre-processed before meaningful markers of stress can be extracted. Data were analyzed using MATLAB vR2016a.

The ECG and PPG data underwent extensive pre-processing to remove baseline wander and a substantial amount of noise, as well as to calculate Inter-Beat Interval (IBI) from the ECG signal and the Peak-to-Peak Interval (PPI) from the PPG data [20]. IBI relates to the time between consecutive R waves (or beats) in an ECG wave, whilst PPI is related to the rate of blood flow, which occurs after a heartbeat. Artifacts in the signals were identified and corrected, including missing peaks and false positives. Details of this process have been described in [20].

The acceleration data from study 1 were also subject to pre-processing, which included first removing the DC offset component in the signal using the notation in (1). Here, A_n is the resulting acceleration vector for each axis (*x*, *y*, and *z*) with the offset removed and *r* is the raw acceleration data for each axis (*x*, *y*, and *z*) that consists of *N* instances.

$$A_n = r - \left(\frac{1}{N}\sum_{i=1}^{N} r_i\right) \quad (1)$$

These data have then been filtered using a 1st order Butterworth lowpass filter, with a cut-off frequency of 30 Hz [21]. Fig. 1 illustrates an example of raw data in the first plot, with the filtered data below in the second plot.

In order to calculate speed, the acceleration signals, which are captured in metres per second squared (m/s^2), must be converted into metres per second (m/s) (i.e. velocity). However, before acceleration can be calculated the separate accelerometer vectors (*x*, *y* and *z*) must be combined into one vector. This function ensures improved accuracy in representing the car's movements (in all directions).

The next stage was to convert the data from m/s^2 to m/s, which achieved using cumulative trapezoidal numerical integration [22] (see (3)), where *v* is the velocity, *t* is the time and *a* is the combined acceleration vector.

$$v_t \simeq \int_t^{a_t} f(a)da \quad (3)$$

The combination of accelerometer vectors was achieved using (2), where the resulting combined accelerometer vector is *A* and the individual accelerometer axis are *Ax*, *Ay*, *Az*.

$$A = \sqrt{Ax^2 + Ay^2 + Az^2} \quad (2)$$

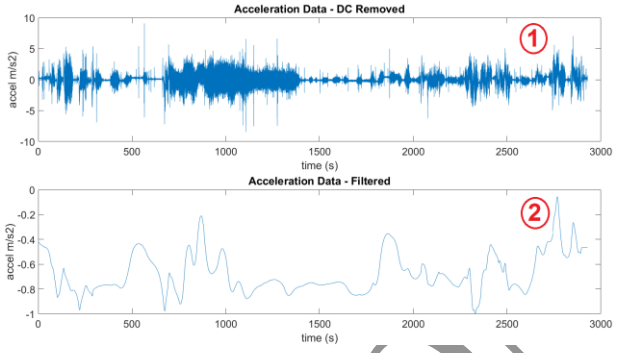

Fig. 1 Example of one of the accelerometer vectors of data with the DC component removed (1) and then filtered (2).

The next item to be calculated was the overall distance covered per drive. An estimation of distance was achieved by calculating the trapezoidal numerical integration of the velocity vector over time [22] (4). Here, *d* is the distance, *t* is the time and *v* is the velocity vector.

$$d_t = \int_{t_1}^{t_2} f(v)dv \quad (4)$$

This resulting value was subsequently divided by the actual distance travelled and multiplied by the velocity (5). In equation 5, *v* is the velocity, *ad* is the actual distance and *d* is the calculated distance from (4).

$$v = o * (ad/d) \quad (5)$$

To verify that *v* has been calculated correctly the trapezoidal numerical integration over time of v has also been calculated (4). The distance value that has been returned will now match the actual distance. Using *v*, we can now extract driving features from the velocity data.

D. Feature Extraction

After pre-processing, the collected data were segmented into 30-second non-overlapping windows and three types of features were extracted – 1) Physiological, 2) Driving and 3) Road.

Physiological features from the processed ECG/PPG signals (study 1 & 2) included calculating heart rate (HR) and heart rate variability (HRV) statistics, which are the predominate factors that are studied for stress detection [23]. HRV is a measure of the standard deviation in inter-beat intervals of successive R waves in a heartbeat [23] and is associated with inflammation [24]. As such, fourteen HRV-related features were extracted from the processed ECG/PPG signals for both studies. These features, per window, included standard time domain features, including mean IBI, heart rate (bpm) and RMSSD, as well as frequency domain features, such as low (0.04-0.14 Hz) and high frequency (0.14-0.40 Hz) bands of HRV.

A number of driving-related features were also extracted for both studies. Study 1 included twenty features, per window, which were extracted from the processed acceleration data. These data included descriptive statistics, such as mean speed, standard deviation and distance travelled, as well as time spent in various speed bands – 0-10 mph, 10-20 mph…>90 mph. Six features, per window, were extracted from the smartphone for study 2, including latitude and longitude coordinates to derive location and speed.

Nine road-related features from the photographs captured in study 2 have been manually extracted. These features pertain to contextual information related to the traffic/road environment and include traffic density (count of moving cars in the lane(s) immediately ahead of the vehicle) and road complexity (count of the number of lanes).

The psychophysiological features were used to undertake heart rate variability (HRV) analysis, which was linked to the driving/road features to provide context of each drive. In this way, we can not only detect stress but determine the contributing factors to those instances of stress. In total, 34 features were extracted for study 1, whilst 29 features were extracted from study 2. The feature sets of each study were then used within the subsequent analysis.

E. Data Labelling

In order to explore the influence of label derivation on classification, the generated feature sets have been labelled using two approaches of 1) subjective questionnaires and 2) physiology, to create ten datasets for analysis (see Fig. 2).

Approach one labelled data via self-reports, which included using the pre- and post-drive responses from the subjective questionnaires to calculate state anger (S-Ang) from the STAXI replies (study 1) and tense arousal (anxious vs. relaxed) from the UMACL questionnaire (study 2). Once pre/post states were calculated, a change score was derived by subtracting the pre-drive score away from the post-drive. Participants who scored < 0 were labelled as relaxed, > 0 were angry and those = 0 were discounted, as no change occurred.

Our second approach labelled data via psychophysiology, which included individually using heart rate (HR), pulse transit time (PTT), high frequency (HF) and low frequency (LF) to create dynamic labels for each participant based on normal distribution. These measures were selected because they are related to various aspects of stress. HR is a measure of activation [25], PTT is implicitly related to blood pressure (BP) [26] and HF/LF are correlated with markers of inflammation [15].

The psychophysiological data were split into percentiles, with data in both the top and bottom 33% of distribution being retained. Stressful labels were assigned to those data that fell into the top 33% of the HR distribution and the bottom percentiles for PTT, HF and LF. Reduced PTT is associated with high blood pressure [27], whereas both HF and LF have been inversely associated with markers of inflammation in the blood [28].

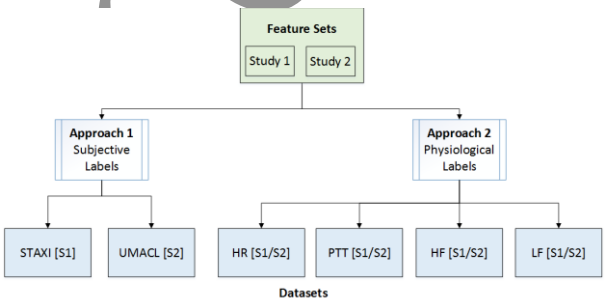

Fig. 2 Process to label the generated feature sets

Instances that fell into the top percentile for PTT, HF and LF were labelled as non-stressful as was HR data in the bottom 33%. Those data points that fell outside the top or bottom 33% of the distribution were not included in the analyses. As depicted in Fig. 2, this process resulted in the creation of ten data sets, which will now be referred to as *STAXI*, *UMACL*, *HR1*, *HR2*, *PTT1*, *PTT2*, *HF1*, *HF2*, *LF1* and *LF2*.

F. Feature Selection

As our approach uses psychophysiology to label data, only the driving and road features are used as the input features within the classification algorithms. Feature selection was performed on each of the ten datasets to reduce the feature spaces and identify those driving/road features that clearly contributed to a discrimination between stressful and non-stressful journeys. The purpose is to identify the relevance of the features that must be must be independent of the input data but cannot be independent of the class [29]. This important stage removes redundant features to reduce the probability of overfitted models. The procedure adopted in this work began by applying Pearson's Correlation Coefficient (6) to the data in order to determine the degree of correlation between driving features. This measurement, *r*, utilizes pairs of features (*x*, *y*) to calculate the linear relation between the features and ranges from -1 to +1 [30].

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \quad (6)$$

Features that produced a correlation score > ± 0.7 with another indicates very strong correlation and so have been removed. For study 1, mean and standard deviation speed were highly correlated with most of the descriptive statistics and so were kept, along with speed bands 0-10 mph…50-60 mph. Features such as distance travelled, median, variance, range, min, max and interquartile range were removed. For study 2, speed and traffic light colour were highly correlated with distance travelled and traffic lights and so were removed as these are measuring similar items. This process reduced the feature space by 55% (study 1) and 40% (study 2).

The second stage involved using the RELIEF algorithm to rank the remaining uncorrelated features. This algorithm uses a nearest-neighbour approach to calculate a distance measure in order to determine the weight of each feature [30]. Features whose weights scored highly can be distinguished among instances that are close to one another [31]. Fig. 3 illustrates an example of this process. Features that occur before the "elbow" of the graph (the point whereby the graph goes from "steep" to "flat") have been kept. This process reduced the uncorrelated feature space by 33% (study 1) and 44% (study 2).

Overall, this feature selection process resulted in a reduction of the feature space by 70% (study 1) and 67% (study 2). The remaining features are now uncorrelated and distinguishable amongst their neighbours. This approach has been repeated independently for each of the datasets depicted in Fig. 2 (*approach 1 – STAXI, UMACL, approach 2 – HR1, HR2, PTT1, PTT2, HF1, HF2, LF1* and *LF2*).

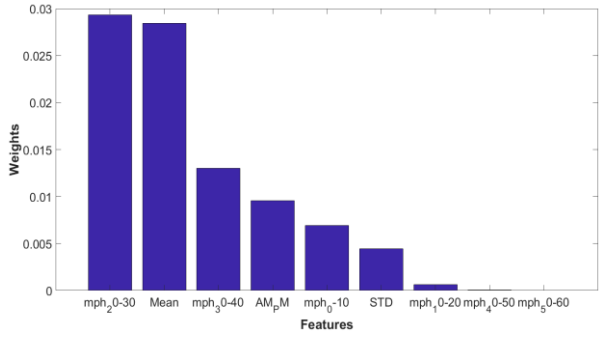

Fig. 3 Example of feature ranking using the RELIEF algorithm on the uncorrelated features in the STAXI dataset

IV. Data Analysis

The analyses involved classifying the reduced datasets using two categories of supervised machine learning algorithms: Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM). These classifiers represent a range of machine learning approaches from high to low bias. The purpose of the evaluation was to demonstrate the accuracy of stress detection based on driving data using: (1) standard method based on high vs. low self-reported levels of anger/anxiety and (2) a self-labelling approach based on upper and lower ranges of psychophysiological reactivity.

The results of the classification were validated using *k*-fold cross-validation, where $k = 10$. Performance measures were calculated for each of the classifier models, including *False Negative Rate* (FNR) – misses that occurs when a stressful/angry drive has been classified as non-stressful/non-angry, *False Positive Rate* (FPR) – false alarms whereby a non-stressful/non-angry drive has been incorrectly classified as stressful/angry, *True Positive Rate* (TPR) [Recall/Sensitivity] – the number of correctly classified stressful/angry drives, *Area Under the Curve* (AUC) – overall performance that measures the tradeoff between the TPR and FPR, and *Balanced Error Rate* (BER) – the average misclassification error rates of each class.

AUC illustrates that the probability of detecting a randomly chosen stressful drive is higher than a randomly chosen relaxed drive. Fig. 4 illustrates the AUC results across both studies. For study 1, the results demonstrated that a combination of SVM and subjective labelled dataset (STAXI1) marginally outperformed the other datasets. However, using both driving/road features and labelling via heart rate (HR2) in study 2 produced significantly better results than subjective labelling.

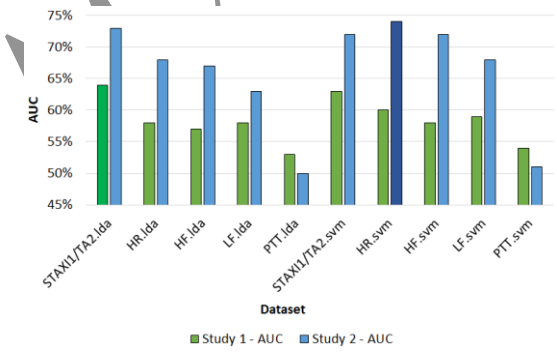

Fig. 4 Area Under the Curve (AUC) results for studies 1 and 2

Study 1

Dataset	FNR	FPR	TPR	BER
STAXI1.lda	38%	42%	62%	40%
HR1.lda	44%	47%	56%	46%
HF1.lda	39%	52%	61%	45%
LF1.lda	45%	47%	55%	46%
PTT1.lda	46%	48%	54%	47%
STAXI1.svm	37%	42%	63%	40%
HR1.svm	38%	47%	62%	42%
HF1.svm	58%	30%	42%	44%
LF1.svm	41%	45%	59%	43%
PTT1.svm	46%	48%	54%	47%

A

Study 2

Dataset	FNR	FPR	TPR	BER
TA2.lda	34%	31%	66%	33%
HR2.lda	32%	35%	68%	34%
HF2.lda	33%	38%	67%	36%
LF2.lda	42%	41%	58%	42%
PTT2.lda	30%	67%	70%	49%
TA2.svm	37%	27%	63%	32%
HR2.svm	34%	31%	66%	33%
HF2.svm	34%	33%	66%	34%
LF2.svm	42%	33%	58%	37%
PTT2.svm	49%	51%	51%	50%

B

Fig. 5 Classification results for a) study 1 and b) study 2

Fig. 5 a) illustrates the results of the classification analyses for study 1. The results demonstrated that combing SVM and subjective labelled dataset (STAXI1) missed the least number of stressful/angry instances and had the lowest misclassification error rate (i.e. low FNR/BER and high TPR). However, it should be noted that a combination of data labelled by HF component of HRV (HF1) and SVM produced the lowest false alarm rate (FPR). Fig. 5 b) illustrates the results of this analysis for study 2. The results indicate that the pulse transit time labelled dataset (PTT2) and LDA achieved better results in terms of 1) low FNR and 2) high TPR. However, the subjective labelled dataset (TA2) and SVM attained equivalent results in terms of low FPR/BER. As such, PTT/LDA missed the least number of stressful/angry instances and was the best performer in terms of correctly classifying stressful/angry drives. Nevertheless, TA/SVM had the lowest false alarm rate (i.e. non-stressful/relaxed drives were incorrectly classified as stressful/angry) and misclassification error.

V. Discussion

This paper demonstrates positive results for detecting negative emotions and cardiovascular stress during real-life driving utilizing dynamically labelled physiological data. Data pertaining to heart rate/HRV and the driving environment were collected via our mobile system across two real-world studies during commuter journeys to/from work. The highest AUC of 74% was achieved using driving data labelled via heart rate. These results illustrate that there is merit in adopting this approach of labelling data via psychophysiology. This is important because using psychophysiological labels, which were defined for each person based on normal distribution, are effectively self-labelling that was *personalised* and *objective* for *each person*. As these labels were derived every 30 seconds, they have much higher fidelity than subjective labels and are sensitive to momentary events during the drive. If we look ahead to driver monitoring systems, and especially systems where feedback is included, the sensitivity of psychophysiological labels to momentary events would be an advantage.

Our work has demonstrated an improvement over similar works. For instance Muaremi et al. [15] collect self-reports four times per day, as well as smartphone data (audio, physical activity and social interaction) and HRV measures. Using a Logistic Regression Model, they achieved accuracies of 55% - 61%. This is comparable to our work whereby dynamically labelling our data via heart rate and utilizing feature selection

has produced improved or similar accuracies without relying on participants to complete multiple self-reports.

Although this work was performed within the context of commuter driving, the approach is extendable. For instance, the system could be extended to detect stress in everyday life and provide real-time feedback. In this instance, it would be beneficial to choose a labelling parameter that exhibits a low false alarm rate, as incorrectly alerting people that they are experiencing stress would be counterintuitive, as it could generate more stress. In this case, utilizing heart rate would be appropriate as this dataset had the lowest rate of false alarms (FPR) of the physiological datasets. However, if the purpose was to monitor people for clinical reasons than failing to detect stress would be worse for people whose health is at risk. In this case, utilizing pulse transit time (blood pressure) would be more appropriate as this dataset had the lowest rate of misses (FNR).

VI. Conclusion and Future Work

Our work has demonstrated a viable method of dynamically labelling data using physiological measures in order to detect periods of negative emotions and cardiovascular stress during real-world driving. This is an important step towards assessing stress "in the wild" without invoking bias from self-reports. Future work aims to investigate the effect of the visualization, from study 2, that participants viewed midway through the collection period.

Acknowledgments

The authors would like to thank all of the participants for agreeing to take part in these studies.

References

[1] Department for Transport, "Transport Statistics: Great Britain," 2015.

[2] T. Triplett, R. Santos, S. Rosenbloom, and B. Tefft, "American Driving Survey 2014–2015," 2016.

[3] D. A. Hennessy and D. L. Wiesenthal, "Traffic Congestion, Driver Stress, and Driver Aggression," *Aggress. Behav.*, vol. 25, no. 6, pp. 409–423, 1999.

[4] M. A. Stults-Kolehmainen, K. Tuit, and R. Sinha, "Lower cumulative stress is associated with better health for physically active adults in the community," *Stress*, vol. 17, no. 2, pp. 157–168, Mar. 2014.

[5] K. Plarre *et al.*, "Continuous Inference of Psychological Stress from Sensory Measurements Collected in the Natural Environment," in *2011 10th International Conference on Information Processing in Sensor Networks (IPSN)*, 2011, pp. 97–108.

[6] World Health Organization (WHO), "Cardiovascular diseases (CVDs): Fact sheet," 2017. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs317/en/. [Accessed: 10-Oct-2017].

[7] B. S. McEwen *et al.*, "Mechanisms of stress in the brain," *Nat. Neurosci.*, vol. 18, no. 10, pp. 1353–1363, Sep. 2015.

[8] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams, "Continuous Stress Detection Using a Wrist Device – In Laboratory and Real Life," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct - UbiComp '16*, 2016, pp. 1185–1193.

[9] J. Healey, "Recording Affect in the Field: Towards Methods and Metrics for Improving Ground Truth Labels," in *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction*, 2011, pp. 107–116.

[10] D. Stokols, R. Novaco, J. Stokols, and J. Campbell, "Traffic Congestion, Type A Behavior, and Stress," *J. Appl. Psychol.*, vol. 63, no. 4, pp. 467–480, 1978.

[11] B. Scott-Parker, C. Jones, and J. Tucker, "Driver stress in response to infrastructure and other road users: simulator research informing an innovative approach to improving road safety," in *Australasian Road Safety Conference*, 2016.

[12] K. Hovsepian, M. Al'Absi, E. Ertin, T. Kamarck, M. Nakajima, and S. Kumar, "Stress: Towards a Gold Standard for Continuous Stress Assessment in the Mobile Environment," in *Proceedings of the ACM International Conference on Ubiquitous Computing (UbiComp)*, 2015, pp. 493–504.

[13] A. Muaremi, A. Bexheti, F. Gravenhorst, B. Arnrich, and G. Troster, "Monitoring the Impact of Stress on the Sleep Patterns of Pilgrims using Wearable Sensors," in *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2014, pp. 185–188.

[14] E. Garcia-Ceja, V. Osmani, and O. Mayora, "Automatic Stress Detection in Working Environments from Smartphones' Accelerometer Data: A First Step," *IEEE J. Biomed. Heal. Informatics*, vol. 20, no. 4, pp. 1053–1060, 2016.

[15] A. Muaremi, B. Arnrich, and G. Tröster, "Towards Measuring Stress with Smartphones and Wearable Devices During Workday and Sleep," *Bionanoscience*, vol. 3, no. 2, pp. 172–183, 2013.

[16] G. Lu, F. Yang, J. A. Taylor, and J. F. Stein, "A comparison of photoplethysmography and ECG recording to analyse heart rate variability in healthy subjects," *J. Med. Eng. Technol.*, vol. 33, no. 8, pp. 634–641, Nov. 2009.

[17] C. D. Spielberger, *The State-Trait Anger Expression Inventory-2 (STAXI-2): Professional Manual*. Odessa, FL: Psychological Assessment Resources, 1999.

[18] P. R. Vagg and C. D. Spielberger, "State Trait Anger Expression Inventory Interpretive Report (STAXI-2: IR)," *Psychol. Assess. Resour. Inc, …*, p. 6, 1979.

[19] G. Matthews, D. M. Jones, and G. A. Chamberlain, "Refining the Measurement of Mood: The UWIST Mood Adjective Checklist," *Br. J. Psychol.*, vol. 81, no. 1, pp. 17–42, 1990.

[20] C. Dobbins and S. Fairclough, "Signal Processing of Multimodal Mobile Lifelogging Data towards Detecting Stress in Real-World Driving," *IEEE Trans. Mob. Comput. (Under Rev.*, 2017.

[21] D. L. Fisher, J. K. Caird, M. Rizzo, and J. D. Lee, "Handbook of Driving Simulation for Engineering, Medicine and Psychology," in *Handbook of Driving Simulation for Engineering, Medicine, and Psychology*, CRC Press, 2011, p. 12.

[22] A. Quarteroni, F. Saleri, and P. Gervasio, *Scientific Computing with MATLAB and Octave*, 2nd ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.

[23] S. Greene, H. Thapliyal, and A. Caban-Holt, "A Survey of Affective Computing for Stress Detection: Evaluating technologies in stress detection for better health.," *IEEE Consumer Electronics Magazine*, vol. 5, no. 4. pp. 44–56, Oct-2016.

[24] V. Papaioannou, I. Pneumatikos, and N. Maglaveras, "Association of heart rate variability and inflammatory response in patients with cardiovascular diseases: Current strengths and limitations," *Front. Physiol.*, vol. 4 JUL, no. July, pp. 1–13, 2013.

[25] P. H. Black and L. D. Garbutt, "Stress, inflammation and cardiovascular disease," *J. Psychosom. Res.*, vol. 52, no. 1, pp. 1–23, Jan. 2002.

[26] G. Sannino, I. De Falco, and G. De Pietro, "Indirect Blood Pressure Evaluation by Means of Genetic Programming," in *Biomedical Engineering Systems and Technologies*, A. Fred, H. Gamboa, and D. Elias, Eds. Springer International Publishing, 2015, pp. 75–92.

[27] D. R. Wagner *et al.*, "Relationship between pulse transit time and blood pressure is impaired in patients with chronic heart failure," *Clin. Res. Cardiol.*, vol. 99, no. 10, pp. 657–664, Oct. 2010.

[28] T. M. Cooper, P. S. McKinley, T. E. Seeman, T. H. Choo, S. Lee, and R. P. Sloan, "Heart rate variability predicts levels of inflammatory markers: Evidence for the vagal anti-inflammatory pathway," *Brain. Behav. Immun.*, vol. 49, pp. 94–100, Oct. 2015.

[29] T. A. Alhaj, M. M. Siraj, A. Zainal, H. T. Elshoush, and F. Elhaj, "Feature Selection Using Information Gain for Improved Structural-Based Alert Correlation," *PLoS One*, vol. 11, no. 11, p. e0166017, Nov. 2016.

[30] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[31] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Boston, MA: Springer US, 1998.