This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2840153, IEEE Transactions on Mobile Computing

Signal Processing of Multimodal Mobile Lifelogging Data towards Detecting Stress in Real-World Driving

Chelsea Dobbins, *Member, IEEE*, Stephen Fairclough

Abstract— Stress is a negative emotion that is part of everyday life. However, frequent episodes or prolonged periods of stress can be detrimental to long-term health. Nevertheless, developing self-awareness is an important aspect of fostering effective ways to self-regulate these experiences. Mobile lifelogging systems provide an ideal platform to support self-regulation of stress by raising awareness of negative emotional states via continuous recording of psychophysiological and behavioural data. However, obtaining meaningful information from large volumes of raw data represents a significant challenge because these data must be accurately quantified and processed before stress can be detected. This work describes a set of algorithms designed to process multiple streams of lifelogging data for stress detection in the context of real world driving. Two data collection exercises have been performed where multimodal data, including raw cardiovascular activity and driving information, were collected from twenty-one people during daily commuter journeys. Our approach enabled us to 1) pre-process raw physiological data to calculate valid measures of heart rate variability, a significant marker of stress, 2) identify/correct artefacts in the raw physiological data and 3) provide a comparison between several classifiers for detecting stress. Results were positive and ensemble classification models provided a maximum accuracy of 86.9% for binary detection of stress in the real-world.

Index Terms— Mobile Computing, Pervasive Computing, Signal Processing, Physiological Measures, Lifelogging, Stress

1 Introduction

Lifelogging is a form of pervasive computing that is concerned with automatically capturing a digital record of an individual's life [1]. This idea was first proposed in 1945 by Vannevar Bush, with the notion of the *Memex* [2]. Since this time, the value of automatically capturing and accessing daily experiences has been appreciated [3]. Earlier work in this domain focused on using wearable cameras to create lifelogging records for self-reflection [4, 5]. However, advancements in technology have enabled a range of sensors to be embedded in smartphones, including cameras, accelerometers, GPS, heart rate sensors, and pedometers, which can be utilized to automatically capture data to supplement lifelogs [6]. Furthermore, the wearable device market is capitalizing on these trends by developing smaller, more powerful and affordable devices that house a multitude of similar sensors.

In order to create truly insightful lifelogs that feed the process of self-reflection, the inclusion of those objective physiological changes that underpin our experiences is vital. As such, leveraging the power of our mobile/wearable devices is essential to access a variety of physiological data, which can be utilized to recognize emotional states [7, 8]. The detection of negative emotions, such as anxiety, stress, sadness and anger, is particularly important as frequent experience of these emotions is associated with inflammatory processes in the cardiovascular system [9]. This process of inflammation may play a significant role in the development of coronary heart disease (CHD) [9, 10]. CHD is the leading cause of death worldwide; however, stress management, via adaptive coping of negative emotions, can reduce the risk of developing CHD [11–13].

Nevertheless, whilst capturing multimodal data from mobile devices is relatively straightforward, the derivation of meaningful information from these sources presents significant challenges. In order to be truly insightful, successful lifelogging systems must integrate multiple streams of data together. This would allow the system to intelligently account for the *context* of physiological measures and their association with the current situation [14]. Context is vital for any lifelogging system and can be defined as "*the state of knowledge of external and internal entities that causes a change in the user's situation, thus necessitating a different interpretation of the data in hand*" [14]. For example, high heart rate correlated to a set of location coordinates and supplemented by a photograph of a red traffic light could indicate an increased physiological response to the experience of journey impedance. In this case, context has been derived from the environment (i.e. from the GPS position and photo), which has then been correlated with the physiological parameters to establish the explanatory framework for the latter. However, the practical achievement of this inferential process is far from straightforward. Collecting and processing covert changes in physiology requires sophisticated digital signal processing techniques and algorithms. Additionally, multiple streams of data (both driving and physiological) must also be synchronized onto a common timeline. This is a significant problem as devices record data at different frequencies.

- *C. Dobbins is with the Department of Computer Science, Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, United Kingdom. E-mail: C.M.Dobbins@ljmu.ac.uk*
- *S. Fairclough is with the School of Natural Sciences and Psychology, Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, United Kingdom. E-mail: S.Fairclough@ljmu.ac.uk*

xxxx-xxxx/0x/$xx.00 © 200x IEEE Published by the IEEE Computer Society

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2840153, IEEE Transactions on Mobile Computing

This work presents our method of processing multimodal lifelogging data to detect stress within the context of real-life driving and forms part of the MultiModal Lifelogging Project (MMLP). This scenario has been chosen because it is a common activity that often includes naturally-occurring episodes of stress. Driving also provides a relatively sedentary and stable environment in which to collect sensor data, as participants remain in a seated position during this activity.

Twenty-one participants took part in two data collection exercises, which required them to collect a variety of lifelogging data on their daily driving commutes to and from work. Their data has been subjected to our data processing pipeline and evaluated using several classification algorithms designed to identify low and high periods of stress. As such, the work addresses the technical challenges of processing a diverse set of signals related to human behaviour on a common time/location basis in order to classify psychophysiological responses.

The remainder of this paper is organized as follows. Section two discusses related work in the area of emotion detection. Section three presents our methodology for preprocessing and extracting features from raw lifelogging data. Section four illustrates the results that have been obtained from classifying our pre-processed data in order to detect stress before providing a discussion of these results in section five. Concluding remarks and directions for future work are then discussed in section six.

2 Related Work

The vision of lifelogging technologies is to, "allow us to capture everything that ever happened to us, to record every event we ever experienced and to save every bit of information we have ever touched" [15]. The sophistication and pervasiveness of mobile and wearable devices has provided an opportunity for this vision to become a reality [16]. Using such devices, a wide range of data can be collected continuously and unobtrusively, enabling the logging of vast amounts of personal data. Extending this area into stress detection via biosensing is an ongoing and exciting research area that promises to deliver increasingly accurate results. Contextual data, such as photos/location, which are typically captured using lifelogging technologies, can be cross-referenced with physiological data in order to identify sources of covert physiological changes.

Measuring stress within drivers usually occurs via simulators [17–19] as there is considerable difficulty, effort and risk involved in collecting data in the natural environment [20]. For instance, Katsis et al. [17] utilized facial electromyography (fEMGs), electrocardiogram (ECG), respiration and skin conductance within support vector machines (SVMs) and adaptive neuro-fuzzy inference system (ANFIS) to detect high stress, low stress, disappointment, and euphoria within a simulated car racing environment. The SVM achieved an overall accuracy of 79%, whilst the ANFIS model achieved 77%. Similarly, Jansen et al. [18] utilised ECG to measure heart rate in order to detect both incidental and integral anger in participants who drove for approximately 12 minutes in a driving simulator. The experience included 9 hazard events (e.g., car swerving into their lane, deer in the road) and afterwards participants rated their affective states using a subjective questionnaire. The results demonstrated that physiological measurements were a valid measurement to use for identifying both incidental and integral affect. However, as these were simulated environments the experimenters could precisely control the road conditions and stability of the sensors.

For the majority of studies who have conducted experiments outside of a laboratory it has been noted that participants often have to follow strict supervision and drive preplanned routes, for a limited time [20]. For instance, Singh et al. [21] have utilised Photoplethysmogram (PPG), Galvanic Skin Response (GSR) and respiration data within a Cascade Forward Neural Network (CASFNN). Data was collected from participants as they drove around three preplanned driving scenarios. The CASFNN achieved an overall accuracy of 80%, using 25 hidden neurons and a 25 second window. However, Vhaduri et al.'s [20] study is similar to this work whereby continuous data has been collected from uncontrolled and unscripted driving episodes over one week. Their work has developed the *GStress* model that estimates driver's stress using only smartphone location (GPS) traces. The model was trained using a Generalized Linear Mixed Model (GLMM) and obtained a Pearson Correlation of 0.722 for predicting stress using only GPS.

Utilizing measures of heart rate variability is an acceptable method to quantify stress [22]. However, coupling these measures with lifelogging technologies can provide insight into those psychological processes, which we may not be consciously aware of. However, in order to advance these fields, conducting experiments outside of the lab and in the field, is an essential step in order to assess the viability of the approach in everyday life.

3 Materials and Methods

Our approach capitalizes on the advancements and availability of smaller and more powerful ambulatory sensors that has enabled us to:

1. Collect instances of raw lifelogging data within real-world driving. These data have been collected from two categories: physiological (wearable) and driving (mobile) sensor data. Physiological data includes raw electrocardiogram (ECG) and photoplethysmogram (PPG). These signals have been used to calculate heart rate, time and frequency-domain measures of heart rate variability (HRV) and pulse transit time (PTT). Driving data includes speed of the vehicle, location, and first-person photographs of the environment.
2. Pre-Process the physiological sensor data to filter noise, calculate various measures, extract features and synchronize with the driving data
3. Detect stress from the synchronized and processed lifelogging data with a high degree of accuracy

However, in order to detect stress, a data processing pipeline is required (see Fig. 1.).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2840153, IEEE Transactions on Mobile Computing

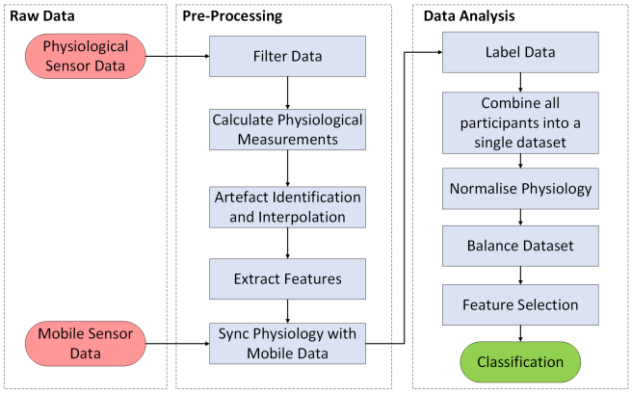

Fig. 1. Data processing pipeline that has been developed to process raw sensor and mobile lifelogging data in order to detect stress.

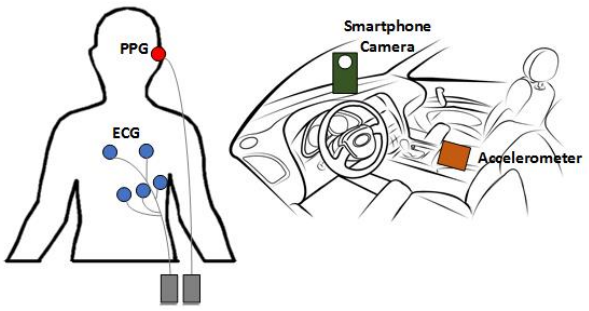

Fig. 2. Subjects wore a Shimmer3 electrocardiogram (ECG) Unit on the chest and clipped a photoplethysmogram (PPG) Optical Pulse Ear-Clip to their ear lobe. An accelerometer was placed in a flat position in the car during DCE A. During DCE B, a smartphone was placed in a holder with the rear camera facing out of the front windshield.

This pipeline has been developed to pre-process and extract features from the collected raw lifelogging data. The remainder of this paper describes this pipeline in more detail.

3.1 Raw Data Collection

Two data collection exercises (DCE) have been undertaken to collect a variety of real-life lifelogging data from participants on their daily driving commutes to and from their place of work.

3.1.1 Participants

The data collection exercises included a total of twenty-one participants – thirteen females and eight males, with an age range from 25 to 57 (mean = 40.86, SD = 11.28). Participants did not have any history of heart disease and were not currently taking any medication that could influence cardiovascular activity. The University Ethical Committee has approved all procedures for participant recruitment and data collection prior to commencement of these studies.

3.1.2 Data Collection Exercise

Raw data was collected using our mobile sensor platform (see Fig. 2) twice a day from participants during their normal driving journeys to and from work, over a period of one week. The protocol included driving for a minimum of 10 minutes (continuously) per journey, driving the same route to/from work at approximately the same time for each journey, being alone in the car (i.e. no passengers) and not listening to music. The journey's ranged from 10:44 minutes to 01:48:30 hours (mean = 34:07 min, SD = 15:52 min).

This mobile sensor platform setup included two wearable Shimmer3™ sensors, which captured both raw electrocardiography (ECG), via a five-lead ECG unit, and photoplethysmogram (PPG) signals, via an optical pulse ear-clip. PPG can be obtained from several areas on the body, including the earlobe and fingertip. The earlobe was chosen because this area provided a stable site for signal collection as opposed to the fingertip, which is highly susceptible to motion artefacts [23], particularly during the driving task.

During DCE A, raw acceleration data was collected via a Shimmer3™ accelerometer unit, which was affixed in a flat position in the car. However, during DCE B the range of driving data that was collected increased to include more contextual information, including photographs, location and speed, which were captured using a custom-built Android application running on a Samsung™ Galaxy S5/S6 smartphone. Photographs were captured every 30 seconds. A mobile phone holder was also provided to place the phone into so that photographs could be taken out of the front windshield.

The Shimmer3™ sensors were configured at a sample rate of 512 Hz. This sampling rate was selected as it was considered to be a suitable frequency at which to obtain a signal that did not suffer from jitter [24]. Data was stored on the internal micro SD card of each device.

Before commencement of the DCE's, participants were briefed and provided with a description of the task and had a demonstration with the equipment. A total of almost 106 hours (525,697,711 instances) of raw lifelogging data have been collected across both DCE's.

3.2 Data Pre-Processing

Collecting lifelogging information produces an extraordinary amount of raw data. In particular, physiological data collected in the field is often susceptible to noise and data loss [25]. For example, the quality of contact that occurs through attaching adhesive electrodes to the skin, can decay over time and even limited physical movement can distort the signal. Therefore, these data must be pre-processed before meaningful markers of stress can be extracted. In the example below, these data were analysed using MATLAB vR2016a.

3.2.1 Filtering

A variety of filtering techniques have been utilized to remove noise and baseline wander. The raw ECG data has been filtered using a Chebyshev Type I second order high pass and lowpass filter, with a cut off frequency between 0.5 Hz and 200 Hz and a passband ripple of 1 dB [26]. The raw PPG data has been filtered using a Chebyshev Type I Lowpass filter, with a passband frequency of 5 Hz and a passband ripple of 1 dB [27]. Once the data were filtered, the next step required heart rate measurements to be calculated from the data.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2840153, IEEE Transactions on Mobile Computing

During DCE A, the acceleration signals were filtered using a Butterworth lowpass filter, with a cut-off frequency of 30 Hz. The signal has also been converted from meters per second squared (m/s^2) into velocity (m/s) using the methods described in [28].

3.2.2 *Calculating Physiological Measurements*

Raw ECG signals record the electrical activity of the heart. The beats of the heart are identified from waves known as the QRS complex [29]. The length of time between consecutive R waves (or beats) is known as the Inter-Beat Interval (IBI). Once a heartbeat occurs, blood flows to different areas of the body and reaches a peak before it progressively decreases [30]. However, a raw PPG signal records the rate of blood flow, which occurs after a heartbeat, as two types of peaks – systolic and diastolic. We were interested in the systolic Peak-to-Peak Interval (PPI), as these are the maximum peaks within the PPG signal. In order to correctly detect stress, accurate detection of the IBI and PPI is essential [31]. In this instance, physiological measurements, including Inter-Beat Interval (IBI) from the ECG signal and the Peak-to-Peak Interval (PPI) from the PPG data, were calculated from the filtered data. However, in order to calculate the IBI and PPI, peaks within both signals must first be detected.

The ECG and PPG data were first segmented using 30-second non-overlapping windows. For each window, the location of the peaks within the ECG and PPG signals were detected. Once the location of the peaks was identified, the IBI and PPI intervals were calculated. This calculation was achieved using the equation in (1). Here, x is the location of the peaks, which was stored as a vector, l is the length of the signal and f is the sample frequency.

$$ibi/ppi = (x(l) - x(l-1)) \div f \times 1000 \quad (1)$$

This equation calculated the difference between adjoining peak locations and then converted this into units of time (milliseconds). Once the IBI and PPI measurements were calculated, the next step required artefacts within the signal to be identified and corrected.

3.2.3 *Artefact Identification and Correction*

When undertaking HRV analysis, artefacts can significantly influence the metrics used to express variability in the heart rate time series [32]. Therefore, it is very important to identify and correct these artefacts. Having a continuous signal is another important issue for HRV analysis hence there is no option to simply discard these artefacts from the record, as this strategy would produce inaccurate metrics [32]. Interpolation is a widely used method to overcome this problem, which corrects artefacts and sustains the integrity of the time series. Our algorithm identified and corrected two types of artefacts, 1) missing peaks and 2) false positives. Fig. 3 illustrates an example of the peaks that have been detected in an ECG signal, the IBI intervals and an example of an identified artefact (this process was also repeated for the PPG signal to generate PPI intervals).

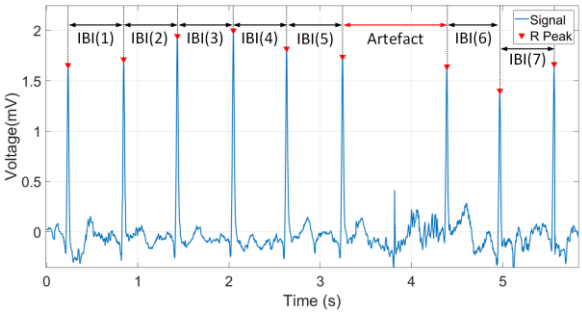

Fig. 3. Example of detected peaks (▼), intervals and artefact in ECG signal.

3.2.3.1 *Identifying Missing Peaks and False Positives*

Algorithm 1 and Fig. 4 presents the process for identifying missing and false positive peaks. The algorithm used the calculated IBI/PPIs from section 3.2.2 (*IBI*) and returned two new binary vectors indicating the position of any 1) missed peaks (*missedPeaks*) and 2) false positive peaks (*fpPeaks*) that have been detected in the windowed signal.

The algorithm looped through each row in the windowed IBI signal (line 1). For each row, if the IBI value was greater than 1.5 of the mean (line 2) this illustrates a significant deviation from normality and so the detection algorithm identifies that a peak has been missed. In this instance, the corresponding row in the *missedPeaks* vector was flagged as 0 (line 3). In the case of identifying false positives, for each row in the windowed IBI signal, if the IBI value was less than 0.5 of the mean (line 7) the detection algorithm identifies that a false positive has occurred. In this instance, the corresponding row in the *fpPeaks* vector was flagged as 0 (line 8). In both instances, if a peak was acceptable then this was flagged with a 1 (lines 5 and 10). Since IBI follows a pronounced normal distribution, these settings were chosen as a method to identify missed peaks and false positives that has been achieved by looking at the deviations from the normal range of values that is specific to each participant during each drive. This process was repeated for the PPG data.

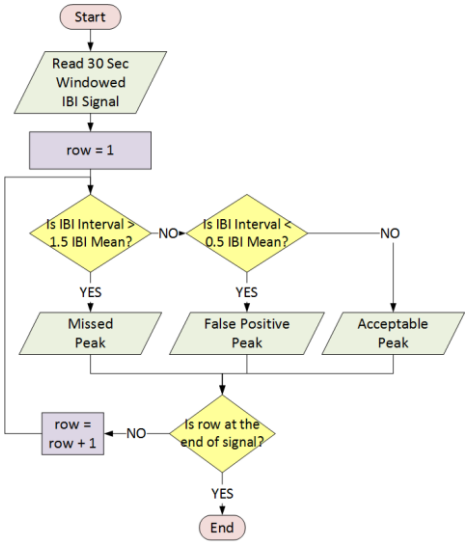

Fig. 4. Flowchart of Algorithm 1 that has been developed to identify missed peaks and false positives

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2840153, IEEE Transactions on Mobile Computing

Algorithm 1. Identify missing peaks and false positives in ECG/PPG signals

Data: *IBI*
Result: *missedPeaks* and *fpPeaks*

1: **for** each row *(j)* in *IBI*
2: **if** *IBI(j)* > (*mean_ IBI* + (*mean_ IBI*/2))
3: *missedPeaks(j)* = 0
4: **else**
5: *missedPeaks(j)* = 1
6: **end if**
7: **if** *IBI(j)* < (*mean_ IBI*/2)
8: *fpPeaks (j)* = 0
9: **else**
10: *fpPeaks (j)* = 1
11: **end if**
12: **end for**
13: return *missedPeaks, fpPeaks*

3.2.3.2 Correcting Missed Peaks and False Positives

Once the missing and false positives peaks had been flagged, Algorithm 2 then corrects these instances by interpolating new peaks and IBI/PPIs (see Fig. 5).

Algorithm 2 uses the flagged *missedPeaks* and *fpPeaks* vectors from Algorithm 1, to obtain all flagged instances that were associated with missing and/or false positives peaks. It then established the number of flagged peaks that occurred and inserted an empty row underneath each flagged instance.

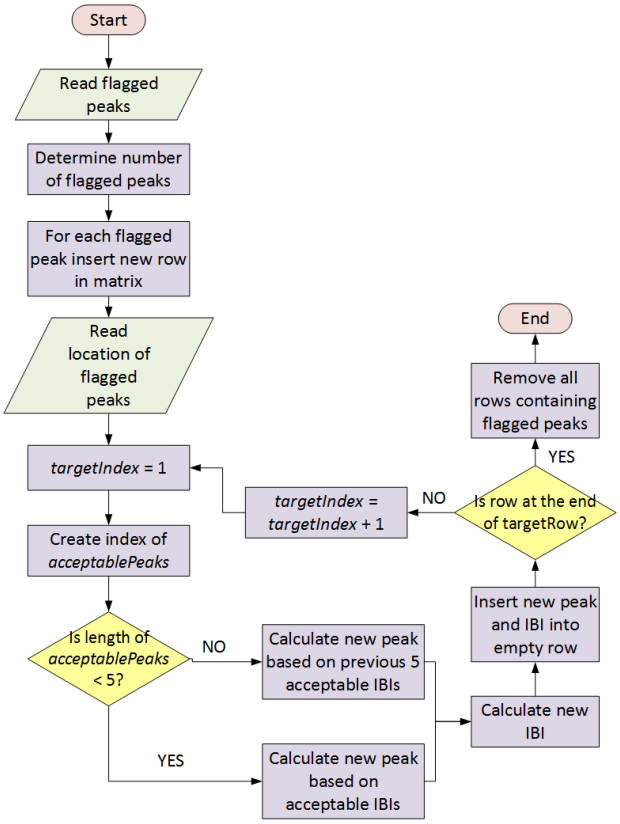

Fig. 5. Algorithm 2 that has been developed to correct missed peaks and false positives

The first item that needed to be corrected were the peaks in the signal. Therefore, the next steps were to get the location of the flagged peaks (*targetIndex*) and loop through each row in the *targetIndex*. For each flagged peak in the *targetIndex*, another index was then created that consisted of the locations of acceptable peaks (*acceptablePeaks*) that occurred *prior* to the flagged peak. A new peak (*np*) was then calculated using equation 2.

$$np = fp - (\bar{x}\ (ibi_n)) \tag{2}$$

This equation uses the flagged peak, *fp*, and the average of the previous five acceptable IBI values, *ibi*, that occurred *before* the flagged peak. However, if the acceptable IBIs occurred at the beginning of the signal and contained less than five values (i.e. *acceptablePeaks* < 5) then *ibi* contained the first *n* < *5* acceptable IBIs that occurred at the start of the signal. In all other instances, *ibi* was based on the previous five acceptable IBIs that occurred prior to the flagged peak. Once the new peak was created, a new corresponding IBI (*nIBI*) value was also created using equation 3.

$$nIBI = np - p_{(fp-1)} \tag{3}$$

This equation uses the newly created peak (*np*) from equation 2 and the previous acceptable peak (*p*) that occurred *before* the flagged peak, *fp*. The new peak (*np*) and corresponding IBI (*nIBI*) were then inserted into the empty row underneath the flagged peak and the flagged IBI was removed. The algorithm terminated once all flagged IBIs in the *targetIndex* were processed and flagged peaks removed.

Using TABLE 1 as an example of this process, a missed peak has been flagged at row 7 and so a new row was inserted underneath (row 8). In order to correct this, a new peak (*np*) was first calculated using equation 2, whereby the average IBI of the previous 5 acceptable IBI's that occurred *before* the missed peak (*cell C2 – C6*) were subtracted from the identified missed peak (*cell B7*) to generate the new peak (*cell B8*).

TABLE 1
EXAMPLE OF CORRECTING MISSED PEAKS AND IBIS IN ECG/PPG SIGNAL

	A	*B*	*C*	*D*	*E*
	R Peak Sample Location	**R Peak Sample Time (ms)**	**IBI (ms)**	**Missed Peak**	**False Positive**
1	121	234.38	0	1	1
2	433	843.75	**609.38**	1	1
3	735	1433.59	**589.84**	1	1
4	1049	2046.88	**613.28**	1	1
5	1348	2630.86	**583.98**	1	1
6	1662	**3244.14**	**613.28**	1	1
7	*2248*	*4388.67*	*1144.53*	*0*	1
8		**3786.72**	**542.58**		

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2840153, IEEE Transactions on Mobile Computing

A new corresponding IBI (*nIBI*) was also created (*cell C8*) by subtracting the previous acceptable peak that occurred *before* the flagged peak (*cell B6*) away from the newly created peak (*cell B8*). Once all flagged items were corrected the flagged rows were removed (i.e. row seven) and so the updated matrix now does not contain any missed peaks and/or false positives. This process occurred for all flagged ECG and PPG peaks. Once the artefacts have been identified and corrected, the next stage involved calculating the Pulse Transit Time and removing any outliers.

3.2.4 Pulse Transit Time and Outlier Removal

Pulse Transit Time (PTT) is indirectly related to blood pressure (BP) and is measured as the time (ms) between an R peak in the ECG and the subsequent S Peak of the PPG signals [33]. As the S Peaks occur after the heartbeat (i.e. ECG) there is a delay, which corresponds to the time it takes for the blood to reach the site of the PPG signal (in our case the earlobe) [30]. However, to get conclusive results, the method relies on these signals being calibrated [27, 33]. Therefore, prior to calculating PTT, the ECG/PPG signals must be inspected for drift, as even the slightest amount of drift within a time window can produce inaccurate data.

Using the processed data from section 3.2.3, Algorithm 3 (see Fig. 6) inspected the PPG signal to determine synchronicity with the ECG signal and returned a matrix of synchronised peaks (*syncPeaks*).

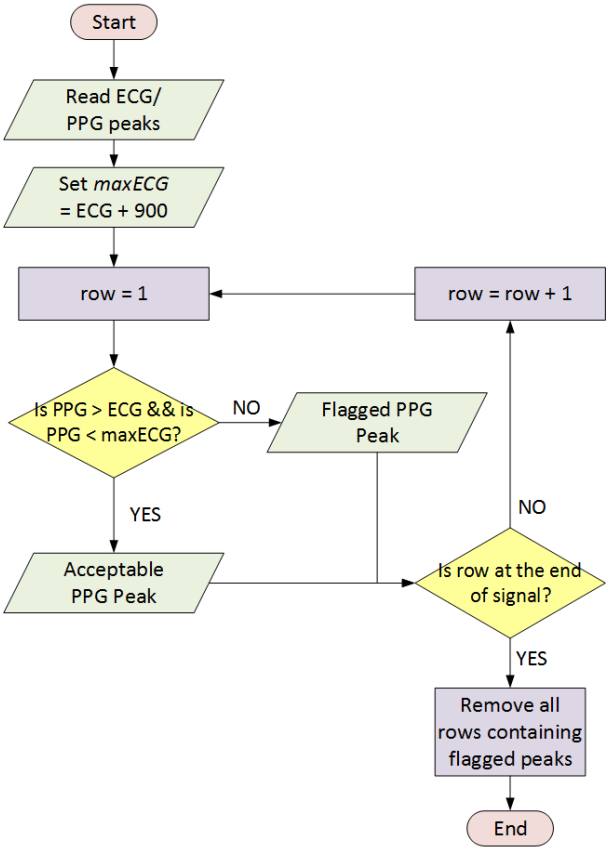

Fig. 6. Flowchart of Algorithm 3 that has been developed to inspect the PPG signal for synchronicity with the ECG

Algorithm 3. Inspect PPG to determine synchronicity with ECG

Data: *R_Peak_Location_ms_ECG, S_Peak_Location_ms_PPG*
Result: *syncPeaks*

1: *maxECG* = *R_Peak_Location_ms_ECG* + 900
2: **for** each row *(j)* in the signal
3: get *rowPPGdata* = *S_Peak_Location_ms_PPG*(*j*)
4: get *rowMaxECG* = *maxECG*(*j*)
5: get *rowECG* = *R_Peak_Location_ms_ECG*(*j*)
6: **if** *rowPPGdata* > *rowECG* && *rowPPGdata* < *rowMaxECG*
7: *corrPPG*(*j*) = 1
8: **else**
9: *corrPPG*(*j*) = 0
10: **end if**
11: **end for**
12: create *syncPeaks* [*R_Peak_Location_ms_ECG, S_Peak_Location_ms_PPG, corrPPG*]
13: remove all rows in *syncPeaks* where *corrPPG* == 0
14: return *syncPeaks*

Using the location of the ECG/PPG peaks as inputs (*R_Peak_Location_ms_ECG* and *S_Peak_Location_ms_PPG*), the algorithm first created a vector (*maxECG*) of the maximum amount of time that should occur between an ECG peak and the subsequent PPG peak (line 1). In this instance, the maximum time should be within 900 ms [34].

For each row in the signal (line 2), the algorithm retrieved the corresponding PPG Peak (*rowPPGdata*), maximum ECG peak time (*maxECG*) and ECG peak (*rowECG*) (lines 3 – 5). If the PPG peak (*rowPPGdata*) was greater than the ECG Peak (*rowECG*) and less than the maximum ECG peak time (*maxECG*) then it was an acceptable PPG peak and the corresponding row in the *corrPPG(j)* vector was flagged as 1 (line 7). However, if the peak was outside of these constraints then the peak was unacceptable and the corresponding row in the *corrPPG(j)* vector was flagged as 0 (line 9). All rows that were flagged as unacceptable (i.e. *corrPPG = 0)* were removed (line 13). The corrected ECG/PPG signals (*syncPeaks*) were then returned (line 14).

PTT was then calculated using equation (4). In this equation, each R Peak ECG sample ($rPeakECG_i$) was subtracted from the corresponding PPG S Peak sample ($sPeakPPG_i$).

$$ptt = sPeakPPG_i - rPeakECG_i \quad (4)$$

The final stage was to use Algorithm 4 to identify outliers within the data (see Fig. 7). Using the calculated PTT data, from equation 4, Algorithm 4 returned a vector of updated PTT values (*PTTupdated*) where any outliers have been removed.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2840153, IEEE Transactions on Mobile Computing

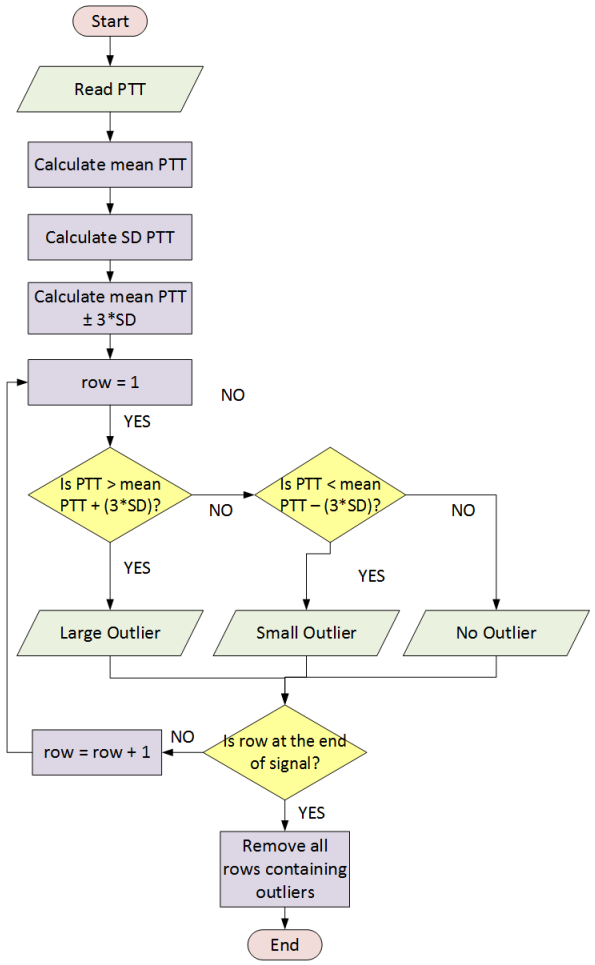

Fig. 7. Flowchart of Algorithm 4 that has been developed to identify and remove outliers from the PTT data

Algorithm 4. Identify and Remove Outliers from PTT

Data: *PTT*
Result: *PTTupdated*

1: calculate mean PTT (*meanPTT*)
2: calculate standard deviation PTT (*stdPTT*)
3:
4: *meanPTTpSD* = *meanPTT* + (3 * *stdPTT*)
5: *meanPTTmSD* = *meanPTT* - (3 * *stdPTT*)
6:
7: **for** each row *(j)* in the PTT signal
8: **if** *PTT(j)* > *meanPTTpSD* || *PTT(j)* < *meanPTTmSD*
9: *largeSmallOut*(j) = 0
10: **else**
11: *largeSmallOut* (j) = 1
12: **end if**
13: **end for**
14: create matrix *PTTupdated* [*PTT*, *largeSmallOut*]
15: remove all rows in *PTTupdated* where *largeSmallOut* == 0
16: return *PTTupdated* vector

In order to identify outliers, the algorithm first calculates the mean (*meanPTT*) and standard deviation (STD) (*stdPTT*) of the PTT data (line 1 – 2). Using these outputs, the mean PTT plus three standard deviations (*meanPTTpSD*) (line 4) and the mean PTT minus three standard deviations were calculated (*meanPTTmSD*) (line 5).

For each row in the PTT vector (line 7), if PTT was greater than *meanPTTpSD* or less than *meanPTTmSD* (line 8) than the corresponding row in the *largeSmallOut(j)* vector was flagged as 0 (line 9), else an outlier was not detected and *largeSmallOut(j)* was flagged as 1 (line 11). All rows that were flagged as outliers (i.e. *largeSmallOut* = 0) were removed (line 15).

To summarise, the developed algorithms in section 3.2.3 have identified and corrected artefacts in the filtered ECG and PPG data, whilst the developed algorithms in section 3.2.4 have calculated pulse transit time (PTT) and have identified and removed outliers. TABLE 2 reports on the number of artefacts that have been identified and removed during this process of artefact correction and outlier removal. The next stage required features to be extracted from this data.

TABLE 2
ARTEFACTS THAT HAVE BEEN IDENTIFIED AND REMOVED FROM THE DATA

DCE	Missed Peaks (%)		False Positives (%)		Large Outliers (%)		Small Outliers (%)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
A	6.2	28.9	0.7	7.0	0.1	0.5	0.1	0.7
B	2.4	6.0	0.2	2.6	0.1	0.6	0.1	0.7

3.2.5 *Feature Extraction*

Using the corrected IBI and PTT signals, several statistical features were extracted from each 30-second non-overlapping window. This is an essential stage as information is difficult to gather from raw data [35].

3.2.5.1 *Physiological Features*

Eleven physiological features have been obtained from the processed IBI and PTT signals in both DCEs. These features included six standard time domain features – Mean IBI, Standard Deviation IBI, Heart Rate, Mean PTT, Standard Deviation PTT and Root Mean Square of the Successive Difference of RR intervals (RMSSD). RMSSD is a measure of parasympathetic heart rate activity, with low values being indicative of reduced parasympathetic activation and high periods of stress [36, 37]. Five features from the frequency domain were also extracted, including:

- Total power (TP) of the signal from 0 – 0.4 Hz
- High frequency (HF) occurring between 0.15 – 0.4 Hz
- Low frequency (LF) occurring between 0.04 – 0.15 Hz
- Very low frequency (VLF) occurring between 0.0033 – 0.04 Hz
- The ratio between low/high frequency (LF/HF)

3.2.5.2 *Driving Features*

Sixteen features related to speed were extracted during DCE A, including driving time (morning/evening), distance travelled (m), mean, median, standard deviation, variance,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2840153, IEEE Transactions on Mobile Computing

range, minimum, maximum and interquartile range of speed (m/s), as well as the time (sec) spent in various speed bands, which ranged from 0-4.5 m/s – 22.4-26.8 m/s.

During DCE B, features extracted from the smartphone included, location (latitude/longitude), speed (m/s), distance travelled (m) and driving time (morning/evening). The photographs have been manually analysed to extract features pertaining to contextual information that are related to the traffic environment, such as traffic density (car count in the lane(s) immediately ahead of the vehicle), road complexity (number of lanes) road type and weather. In total, twelve driving features have been extracted from the smartphone.

In total, twenty-seven features have been extracted during DCE A, whilst twenty-three features have been extracted during DCE B. The physiological, photograph and driving features from DCE B were amalgamated into one matrix on a common time basis of 30 second windows. Location data (i.e. latitude/longitude coordinates) were also matched and appended to each time window.

4 Data Analysis

4.1 Data Labelling

Questionnaires were used to capture the subjective changes in mood that occurred due to each journey. DCE A utilized a short-version of the State–Trait Anger Expression Inventory 2 (STAXI 2) [38] questionnaire, which was composed of fifteen statements (e.g. I am furious, I feel like yelling at somebody, etc.). Participants had to score their current feeling in relation to each statement on a Likert scale, whereby 1 = *not at all*, 2 = *somewhat*, 3 = *moderately so* and 4 = *very much so*. However, it was noted that social desirability may have influenced the responses as there seemed to be a reluctance to admit negative feelings.

In response to this issue, a short-version of the UWIST Mood Adjective Checklist (UMACL), which has been developed and validated by Matthews et al [39], was used instead during DCE B. The questionnaire is composed of fourteen words that described feelings (e.g. happy, relaxed, sad, angry, etc.). Participants were required to rate how well each word described their current mood state on a Likert scale, where 1 = *definitely*, 2= *slightly*, 3 = *slightly not* and 4 = *definitely not*.

Both questionnaires were administered using a custom-made Android application and were completed *before* and *after* each journey to account for any changes in mood that occurred during the duration of the drive. The scores from the subjective questionnaires were processed to derive a change score (post-drive – pre-drive). Change scores related to the feeling of negative emotions were used as subjective labels for the data to describe the level of stress associated with each journey. Those journeys that scored a change score a) above zero were labelled as *stressful*, b) below zero were labelled as *non-stressful* and c) equal to zero were *discounted* as a change was not noted. Fig. 8 illustrates the frequency of journeys for each category.

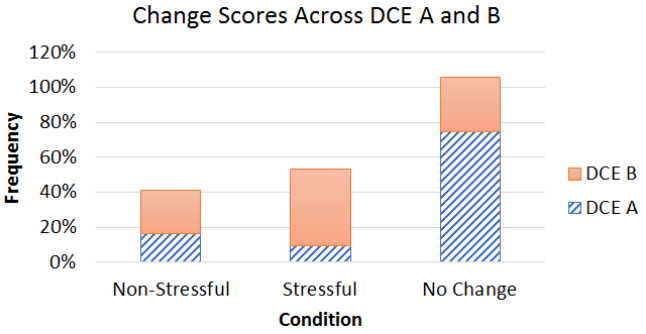

Fig. 8. Change scores across DCE A and B

For each DCE, data pertaining to each drive/participant were amalgamated and physiological data were normalized by calculating the z-score of each feature to account for individual differences between participants. These two labelled datasets (DCE A and DCE B) formed the basis for our analysis into detecting stress from multimodal lifelogging data.

However, as the datasets are unbalanced, it was necessary to balance the minority class before the analysis could occur. The Synthetic Minority Over-Sampling Technique (SMOTE) has been used to generate new synthetic records to balance the dataset. This approach is an accepted technique for solving the problems related to unbalanced datasets [40].

4.2 Feature Selection

Feature selection was performed to reduce the datasets into a subset of those features that clearly contributed to a discrimination between stressful and non-stressful journeys. However, the analysis involved utilizing a number of supervised machine learning algorithms to classify the data using a) only driving features, b) only physiological features and c) an amalgamation of a and b (i.e. both driving and physiological features were merged together into one dataset of features). The purpose of this was to investigate the most appropriate type of features to use for detecting stress. As such, the process of feature selection was undertaken separately on both types of features to select the best driving and physiological features, on each dataset.

In order to remove irrelevant attributes features were ranked using the RELIEFF algorithm [41]. This algorithm uses a k nearest neighbour approach to find the average contribution of all k nearest hits and misses. This average is then weighted with the prior probability of each class to estimate the quality of the features. The ranked weights and features were plotted and eliminated based on the "elbow" of the graph, the point whereby the graph goes from "steep" to "flat". Fig. 9 illustrates an example of a graph that has been plotted for DCE A's driving features.

TABLE 3 illustrates the features that have emerged as the top ranked variables that distinguished Stressful from Non-Stressful journeys within DCE A and B's data. This analysis has removed 69% and 58% of the driving features and 55% and 27% of the physiological features from DCE A and B's datasets (respectively).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2840153, IEEE Transactions on Mobile Computing

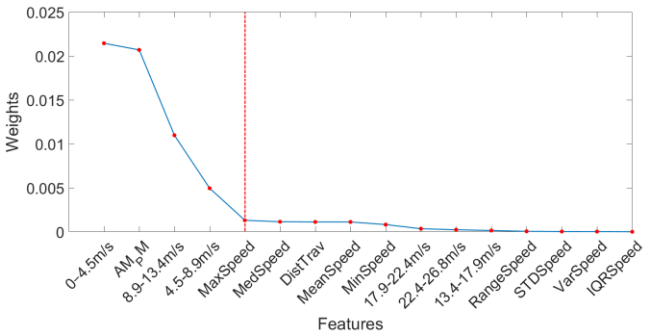

Fig. 9. Example of RELIEFF feature selection. Features that occur after the "elbow" of the graph have been removed.

TABLE 3
TOP RANKED FEATURES THAT HAVE BEEN SELECTED FOR EACH DATASET

Driving Feature	Weight	Physiological Feature	Weight
DCE A			
0 – 4.5 m/s	0.0214	Mean PTT	0.0063
AM_PM	0.0207	HR	0.0046
8.9 – 13.4 m/s	0.0110	STD PTT	0.0020
4.5 – 8.9 m/s	0.0050	LF_HF	0.0015
Max Speed	0.0013	HF	0.0009
DCE B			
Time Day	0.1246	Mean PTT	0.0264
AM_PM	0.0995	Mean IBI	0.0172
In Traffic	0.0501	STD PTT	0.0166
Distance Travelled	0.0266	RMSSD	0.0161
Car Count	0.0241	STD IBI	0.0150
		HF	0.0121
		TP	0.0084
		LF	0.0038

The features identified in TABLE 3 were then used within the subsequent evaluation.

4.3 Classifier Performance

The evaluation is based on a user-independent model that utilized both parametric and non-parametric classifiers, including Linear Discriminant Analysis (LDA), Decision Tree (DT) and *k*-Nearest Neighbours (kNN), to differentiate between stressful and non-stressful journeys. An ensemble classifier was also built, which weighted and combined the predictions of the above classifiers using the Hill-Climbing algorithm [42, 43]. The benefit of using an ensemble approach is that bias, variance and overfitting are reduced.

Each classifier and the ensemble approach were evaluated independently using a) only the driving features, b) only the physiological features and c) an amalgamation of a and b (i.e. both driving and physiological features were merged together into one dataset of features). Fig. 10 illustrates the approach that has been used for the classification analysis.

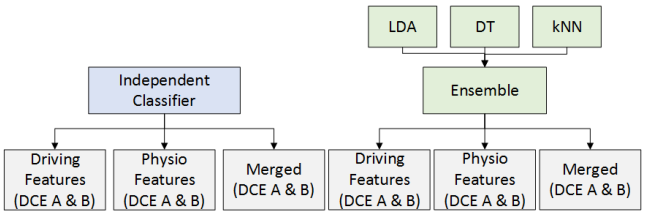

Fig. 10. Classification approach that has been used during the evaluation.

The results were validated using repeated *k*-fold cross-validation, whereby k = 10 and repetitions = 100. The performance measurements that were calculated included:

- *Accuracy* – An index of overall performance
- *F_1 Score* – The harmonic mean of *Precision* [*Positive Predictive Value*] – the proportion of results that have been marked as positive (stressful) where a true positive (stress) has actually occurred and *Recall* [*True Positive Rate/Sensitivity*] – the proportion of stressful drives (positives) that are correctly identified as being stressful (positive).
- *Balanced Error Rate (BER)* – The average errors of each class.
- *Receiver Operating Characteristics (ROC) Curve* – Summary of performance that plots the *True Positive Rate (TPR) [Recall/Sensitivity]* against the *False Positive Rate (FPR) [Type I Error]* – false alarms that indicates that an instance has been classified as stressful when stress is actually not present.

TABLE 4 illustrates that during DCE A, the independent classifiers LDA and DT produced comparable accuracies to the ensemble approach (61.33%, 61.06% and 61.29% respectively) and error rates (38.61%, 38.91% and 38.61% respectively). This pattern demonstrates that these classifiers were similar in their performance of detecting stressful journeys and in the amount of errors that were produced for each class. However, DT outperformed the others and had the highest F1 (65.43%), which illustrates that there was a higher balance between precision and recall, i.e. correctly detecting a stressful drive when stress has actually occurred. This illustrates that for features related only to speed a simple linear model will suffice. However, during DCE B, the ensemble approach outperformed the independent classifiers in terms of the highest accuracy, F1 and lowest BER. This pattern demonstrates that when contextual data is introduced, in addition to speed, and the classifiers are combined the results improve. Overall, the ensemble approach in conjunction with contextual features about the drive achieved the best performance across both DCEs.

TABLE 5 illustrates that using only physiological features improved upon the driving features. Furthermore, during both DCEs, the ensemble approach outperformed the independent classifiers in terms of higher accuracy (65.04% and 80.04%), F1 (66.3% and 78.98%) and lower BER (34.96% and 19.91%).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2840153, IEEE Transactions on Mobile Computing

TABLE 4
CLASSIFIER PERFORMANCE FOR DRIVING FEATURES ONLY

	DCE A				DCE B			
Measurement	***LDA***	***DT***	***kNN***	***Ensemble***	***LDA***	***DT***	***kNN***	***Ensemble***
Accuracy	**61.33%**	61.06%	58.11%	61.29%	74.92%	75.31%	75.27%	**77.28%**
F_1	59.32%	**65.43%**	57.78%	62.80%	74.69%	75.64%	74.64%	**76.92%**
BER	**38.61%**	38.91%	41.88%	**38.61%**	25.07%	24.68%	24.73%	**22.72%**

TABLE 5
CLASSIFIER PERFORMANCE FOR PHYSIOLOGICAL FEATURES ONLY

	DCE A				DCE B			
Measurement	***LDA***	***DT***	***kNN***	***Ensemble***	***LDA***	***DT***	***kNN***	***Ensemble***
Accuracy	58.96%	62.75%	63.72%	**65.04%**	73.16%	75.66%	78.65%	**80.04%**
F_1	59.37%	65.66%	64.59%	**66.30%**	70.86%	75.47%	76.49%	**78.98%**
BER	40.96%	37.33%	36.29%	**34.96%**	27.02%	24.02%	21.60%	**19.91%**

TABLE 6
CLASSIFIER PERFORMANCE FOR MERGED DRIVING AND PHYSIOLOGICAL FEATURES

	DCE A				DCE B			
Measurement	***LDA***	***DT***	***kNN***	***Ensemble***	***LDA***	***DT***	***kNN***	***Ensemble***
Accuracy	63.29%	64.29%	69.26%	**69.73%**	81.73%	78.86%	86.02%	**86.86%**
F_1	64.06%	67.44%	69.72%	**70.40%**	80.42%	78.05%	84.72%	**85.89%**
BER	36.69%	35.82%	30.75%	**30.27%**	18.31%	20.92%	14.12%	**13.16%**

This illustrates that the overall performance and quality were greatly improved and a high level of balance between precision and recall, as well as a lower error rate, was produced when the independent models were combined.

TABLE 6 demonstrated the best results, which occurred when both driving and physiological features were amalgamated into one dataset and used in conjunction with ensemble learning. This approach generated the highest overall accuracy (86.86%), F1 (85.89%) and lowest BER (13.16%) across TABLE 4, TABLE 5 and TABLE 6.

ROC curves have been produced to summarise the performance of the ensemble classification method for each set of features (see Fig. 11). As it can be seen in Fig. 11, merging both driving and physiological features into one dataset produces a high probability of detecting that a stressful drive will be correctly identified when stress was present, whilst ensuring that falsely classifying an instance as stressful when stress is not present is minimized.

To summarise, the results confirm the conclusions that may be drawn from these results, which illustrates that using both driving and physiological features, in conjunction with ensemble learning, may be the most appropriate classifier for the detection of stress.

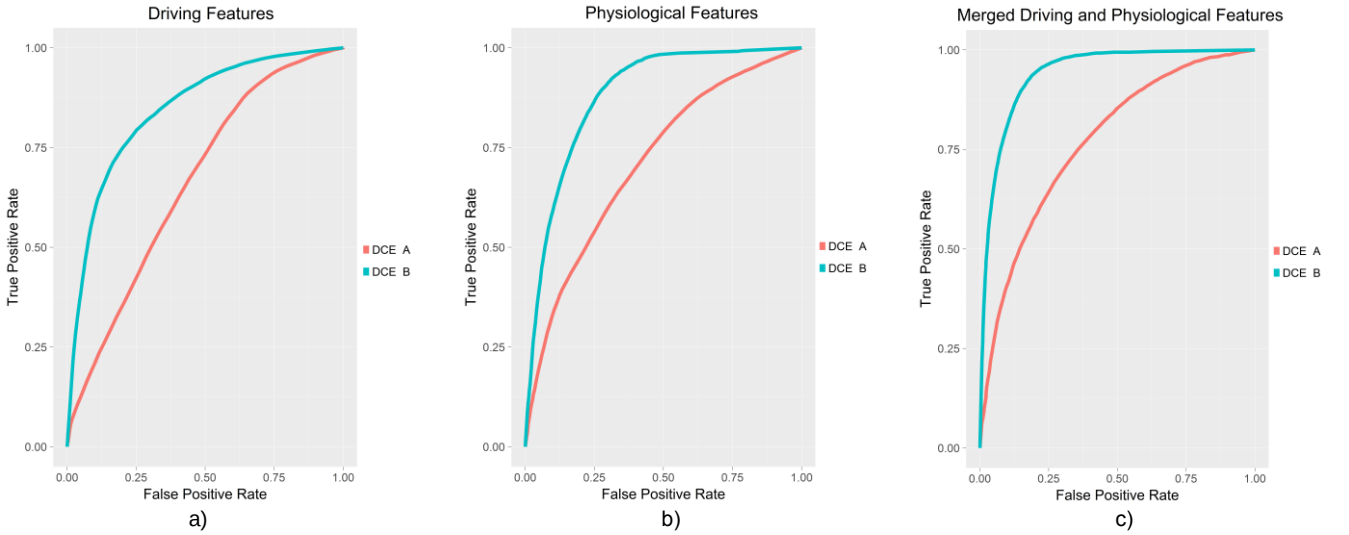

Fig. 11. ROC Curves of the Ensemble classification approach for DCE A and DCE B using a) driving features b) physiological features and c) an amalgamation of driving and physiological features

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2840153, IEEE Transactions on Mobile Computing

5 Discussion

This paper demonstrates the feasibility of applying our signal processing approach to real-world multimodal lifelogging data. These data were collected using mobile/wearable devices during everyday driving with the aim of detecting those journeys that were associated with increased stress.

In order to demonstrate the feasibility of our approach, we performed two data collection exercises (DCE A & B). The first piece of data collection relied exclusively on speed data to characterise the driving environment and data were labelled on the basis of responses to the STAXI questionnaire, which specifically captures the subjective experience of anger. This experience led to two key developments of our experimental protocol for DCE B. In the first instance, we coded events captured on the camera to increase the range of variables obtained from the driving environment, e.g. number of vehicles, weather, road type. In addition, we switched from the STAXI to the UMACL, which is a questionnaire designed to index subjective mood. This latter decision represented a response to the shortcomings of the STAXI questionnaire. It was apparent during the first data collection exercise that responses to the STAXI was influenced by social desirability. Many participants were either reluctant to acknowledge increased anger or their experience of anger was transitory and had disappeared when the journey was over. This trend is apparent in Fig. 8 by the number of subjective responses where no change was observed. The UMACL, on the other hand, takes the form of a mood adjective checklist, which is a less direct method of assessment than STAXI and shifts the emphasis towards feelings of tension, which are more socially acceptable than an expression of anger. The choice of self-report tool is particularly important for this type of evaluation, where labels for classification are derived from subjective self-assessment. It is important that any subjective questionnaire that is incorporated into this type of investigation is capable of quantifying self-reported states with a high degree of accuracy and sensitivity.

Our approach to classification involved a number of distinct phases that were designed in order to gauge the relative contribution of variables derived from driving and physiology. The application of the RELIEFF algorithm (TABLE 3) demonstrated that driving features that captured episodes of journey impedance (e.g. slow speed, high car count) were well represented, as was time of day. With respect to the latter, we would conclude that traffic density was higher in the late afternoon compared to the morning, hence variables related to time of day were effectively proxies for journey impedance. It was noted that PTT was the physiological feature with the highest score for both data sets, presumably due to its association with blood pressure. Heart rate and measures related to heart rate variability were also selected, particularly high frequency of heart rate variability (HRV), which is associated with parasympathetic activation and inflammation.

The methodology for classification was designed to test both driving and physiological data from both data sets using a range of algorithms both alone and as an ensemble (TABLE 4 – TABLE 6). With respect to driving data and using F1 as a performance indicator, there was little differentiation between the three algorithms for DCE B, whereas Decision Trees (DT) showed a significant advantage for DCE A (TABLE 4). As a general trend for classification using driving data, particularly when looking at ensemble performance, DCE B performed substantially higher (76.92%) compared to DCE A (62.8%). We assume this advantage was achieved by extending the range and variety of driving variables in DCE B beyond those measures of speed used in DCE A. If we consider the results of classification using physiological data (TABLE 5), once again using F1 as a measure of performance, it is noted that both DT and kNN models deliver superior classification to LDA. A comparison of ensemble performance shows a clear advantage for DCE B (78.98%) over DCE A (66.3%), presumably due to the higher number of physiological features selected by the RELIEFF algorithm during the feature selection phase (TABLE 3).

Those subjective states experienced by the driver during a commuter journey, whether they are associated with anger or anxiety, represent an amalgamation of the driving environment and the physiological responses of the individual to that driving environment. This is the reason why those classification models that merged both sets of features delivered higher classification accuracy compared to those based on either driving or physiology alone (TABLE 6). If we look at ensemble performance (using F1) for DCE A, we see classification performance of 70.4% (TABLE 6) compared to 62.8% (driving) and 66.3% (physiology) from the equivalent models in TABLE 4 and TABLE 5. The same trend was observed for DCE B where ensemble classification was 85.89% (TABLE 6) compared to 76.92% and 78.98% for driving (TABLE 4) and physiology (TABLE 5) respectively. The use of physiological features for classification of psychological states in the real world is significantly enhanced by the inclusion of features related to the context of those psychological states.

The availability and miniaturization of sensors has enabled the continuous measurement of quantifiable data in everyday life. However, as observed by Hovsepian et al. [25], we are still lacking a well-validated stress model that can be used for managing stress in the natural environment. For a model to be considered a “gold standard” for continuous stress assessment, a high accuracy of ≥ 70% outside a lab setting (in the field) is required [25]. The results from this study are positive and provide a successful method of pre-processing mobile lifelogging physiological and driving sensor data to achieve a maximum accuracy of 86.9% in detecting stress (TABLE 6).

Our work demonstrated an improvement over similar works in the area of detecting stress “in the wild”. For instance, Hovsepian et al. [25] utilized ECG, HRV and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2840153, IEEE Transactions on Mobile Computing

respiration features within a support vector machine (SVM) to classify stress. Their data has been labelled using self-reports of stress that have been obtained using an adaptation of the Perceived Stress Scale (PSS). Their results demonstrate an accuracy of 72% in the field. However, our work has improved upon this by achieving a maximum accuracy of 86.9% (TABLE 6), which could be attributed to the method that has been applied to pre-process our data and the selection of features that has been used. This work [25] utilized 37 features, whereas in our work we have reduced our feature set to five using feature selection to select a subset of those features that effectively discriminated stressful drives from non-stressful ones. Most importantly, we have utilized primarily HRV-related features, including RMSSD, which can be calculated in real-time and is correlated with markers of inflammation [22]; for critical assessment of this link, see [44].

The collection of ambulatory data outside of a laboratory presented a number of challenges, such as data loss (due to physical artefacts), a reliance on participants operating the sensors properly and completing the data collection protocol consistently and correctly. Although laboratory experiments offer greater control over experimental variables, they suffer with respect to ecological validity of the phenomenon under investigation [45]. The presence of potential confounds and loss of control over the environment that characterizes work in the field is the price to be paid for taking research on stress out of the laboratory. This transition can also inform the development and testing of mobile applications as their usability can only be properly evaluated in the field [45]. Furthermore, as discussed in previous work [46], lifelogging research tends to lack robust data analytical approaches and real-world datasets. As such, there is a pressing need to develop validated approaches to pre-processing real-world data so that such applications can be taken forward for use in that research community. The novelties of the work that we have described include:

1) Providing a set of algorithms for pre-processing raw lifelogging data that has been obtained from mobile/wearable devices in order to calculate valid measures of heart rate variability
2) Providing a set of algorithms for artefact identification and interpolation so that missing peaks and false positives can be corrected
3) Providing a comparison between several classifiers to determine the most appropriate approach for detecting stress. The accuracy of the stress detection is significantly enhanced when features related to the physiology and context are included in the classification task.

This work also has implications for advancing the field of lifelogging. By combining traditional lifelogging techniques with psychophysiological signals to quantify negative states and their physiological correlates, which we may not be overtly aware of, can deliver a greater understanding of environmental triggers for those negative states. This benefit may have implications for long-term health as the repeated experience of stress can induce a chronic inflammatory process that can culminate in atherosclerosis (a build-up of fatty material inside arteries that makes a major contribution to heart attacks/strokes) [47].

6 Conclusions and Future Work

Our work demonstrated a viable method of pre-processing raw lifelogging data in order to calculate valid measures of heart rate variability and correct artefacts for the purpose of classifying periods of stress during real-world driving.

Our approach has also provided an improvement over the level of accuracy achieved in comparison to other works in the area of detecting stress “in the wild”. Nevertheless, there are limitations in the study that could be improved upon via further investigation. For instance, this work has labelled data based on the results of the subjective questionnaires that were captured before/after each drive, however this approach has significant limitations for labelling psychophysiological data and measures from the driving environment, both of which fluctuates in real-time. In addition to subjective self-report data being associated with retrospective bias and having limited fidelity, questionnaire data can only represent the conscious experience of the individual, whereas psychophysiological data responds to both conscious and subconscious processes. An interesting line of enquiry would be to label the data based on either psychophysiology or driving conditions and compare those results with the subjective labels. Labelling via physiology/driving conditions would overcome those limitations associated with self-reporting. Additionally, exploring user-dependent models is another line of enquiry that is worth pursuing in order to build models that can be personalised to the individual. Further research is required to explore these ideas and to assess if the findings can be replicated in other domains of emotion detection.

Acknowledgment

This work has been supported by the UK Engineering and Physical Sciences Research Council (EPSRC) under Research Grant EP/M029484/1. Owing to ethical concerns/the sensitive nature of this research, the data underlying this publication cannot be made openly available. Further information, including conditions for metadata access, can be found at LJMU Data Repository, at http://opendata.ljmu.ac.uk/24/. Dr Chelsea Dobbins is the corresponding author of this paper.

References

[1] A. L. Allen, “Dredging Up the Past: Lifelogging, Memory and Surveillance,” *Univ. Chicago Law Rev.*, vol. 75, pp. 47–74, Sep. 2008.

[2] V. Bush, “As We May Think,” *The Atlantic Monthly*, no.

JULY 1945, 1945.

[3] M. M. Stevens, G. D. Abowd, K. N. Truong, and F. Vollmer, "Getting into the Living Memory Box: Family archives & holistic design," *Pers. Ubiquitous Comput.*, vol. 7, no. 3–4, pp. 210–216, Jul. 2003.

[4] A. R. Doherty *et al.*, "Experiences of Aiding Autobiographical Memory Using the SenseCam," *Human–Computer Interact.*, vol. 27, no. 1–2, pp. 151–174, 2012.

[5] S. Hodges *et al.*, "SenseCam: A Retrospective Memory Aid," *UbiComp 2006 Ubiquitous Comput.*, vol. 4206, pp. 177–193, 2006.

[6] R. Francese, M. Risi, G. Tortora, and M. Tucci, "Visual Mobile Computing for Mobile End-Users," *IEEE Trans. Mob. Comput.*, vol. 15, no. 4, pp. 1033–1046, Apr. 2016.

[7] L. Ivonin, H.-M. Chang, W. Chen, and M. Rauterberg, "Unconscious emotions: quantifying and logging something we are not aware of," *Pers. Ubiquitous Comput.*, vol. 17, no. 4, pp. 663–673, Apr. 2012.

[8] D. McDuff, A. Karlson, A. Kapoor, A. Roseway, and M. Czerwinski, "AffectAura: An Intelligent System for Emotional Memory," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, 2012, pp. 849–858.

[9] J. K. Kiecolt-Glaser, L. McGuire, T. F. Robles, and R. Glaser, "Emotions, Morbidity, and Morality: New Perspectives from Psychoneuroimmunology," *Annu. Rev. Psychol.*, vol. 53, no. 1, p. 83, Feb. 2002.

[10] J. Suls and J. Bunde, "Anger, Anxiety, and Depression as Risk Factors for Cardiovascular Disease: The Problems and Implications of Overlapping Affective Dispositions," *Psychol. Bull.*, vol. 131, no. 2, pp. 260–300, Mar. 2005.

[11] NHS, "Coronary heart disease," 2014. [Online]. Available: http://www.nhs.uk/conditions/Coronary-heart-disease/Pages/Introduction.aspx. [Accessed: 20-Oct-2016].

[12] U.S. Department of Health and Human Services, "How Can Coronary Heart Disease Be Prevented or Delayed?," 2016. [Online]. Available: http://www.nhlbi.nih.gov/health/health-topics/topics/cad/prevention. [Accessed: 20-Oct-2016].

[13] U.S. Department of Health and Human Services, "Managing stress," 2016. [Online]. Available: http://www.nhlbi.nih.gov/health/health-topics/topics/heart-healthy-lifestyle-changes/managing-stress. [Accessed: 20-Oct-2016].

[14] H. Viswanathan, B. Chen, and D. Pompili, "Research Challenges in Computation, Communication, and Context Awareness for Ubiquitous Healthcare," *IEEE Commun. Mag.*, vol. 50, no. 5, pp. 92–99, May 2012.

[15] A. Sellen, A. Fogg, M. Aitken, S. Hodges, C. Rother, and K. Wood, "Do Life-Logging Technologies Support Memory for the Past? An Experimental Study Using SenseCam," in *Conference on Human Factors in Computing Systems, CHI '07, Irvine, CA*, 2007, pp. 81–90.

[16] R. Rawassizadeh, M. Tomitsch, K. Wac, and A. M. Tjoa, "UbiqLog: a generic mobile phone-based life-log framework," *Pers. Ubiquitous Comput.*, vol. 17, no. 4, pp. 621–637, Apr. 2012.

[17] C. D. Katsis, N. Katertsidis, G. Ganiatsas, and D. I. Fotiadis, "Toward Emotion Recognition in Car-Racing Drivers: A Biosignal Processing Approach," *IEEE Trans. Syst. Man, Cybern. - Part ASystems Humans*, vol. 38, no. 3, pp. 502–512, May 2008.

[18] S. Jansen, A. Westphal, M. Jeon, and A. Riener, "Detection of Drivers' Incidental and Integral Affect Using Physiological Measures," in *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '13)*, 2013, pp. 97–98.

[19] D. MacLean, A. Roseway, and M. Czerwinski, "MoodWings: A Wearable Biofeedback Device for Real-Time Stress Intervention," in *Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '13)*, 2013, pp. 1–8.

[20] S. Vhaduri, A. Ali, M. Sharmin, K. Hovsepian, and S. Kumar, "Estimating Drivers' Stress from GPS Traces," in *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '14)*, 2014, pp. 1–8.

[21] R. R. Singh, S. Conjeti, and R. Banerjee, "Assessment of Driver Stress from Physiological Signals collected under Real-Time Semi-Urban Driving Scenarios," *Int. J. Comput. Intell. Syst.*, vol. 7, no. 5, pp. 909–923, Sep. 2014.

[22] 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.

[23] 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.

[24] R. J. Ellis, B. Zhu, J. Koenig, J. F. Thayer, and Y. Wang, "A careful look at ECG sampling frequency and R-peak interpolation on short-term measures of heart rate variability.," *Physiol. Meas.*, vol. 36, no. 9, pp. 1827–1852, Sep. 2015.

[25] 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.

[26] S. Rani, A. Kaur, and J. S. Ubhi, "Comparative study of FIR and IIR filters for the removal of Baseline noises from ECG signal," *Int. J. Comput. Sci. Inf. Technol.*, vol. 2, no. 3, pp. 1105–1108, 2011.

[27] K. Li and S. Warren, "Initial Study on Pulse Wave Velocity Acquired From One Hand Using Two Synchronized Wireless Reflectance Pulse Oximeters," in *2011 Annual International Conference of the IEEE Engineering in*

Medicine and Biology Society (EMBC), 2011, pp. 6907–6910.

[28] C. Dobbins and S. 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) [Accepted]*, 2018.

[29] Z.-E. Hadj Slimane and A. Naït-Ali, "QRS complex detection using Empirical Mode Decomposition," *Digit. Signal Process.*, vol. 20, no. 4, pp. 1221–1228, Jul. 2010.

[30] V. Kalidas and L. S. Tamil, "Cardiac arrhythmia classification using multi-modal signal analysis," *Physiol. Meas.*, vol. 37, no. 8, pp. 1253–1272, Aug. 2016.

[31] M. U. Ahmed, S. Begum, and M. S. Islam, "Heart Rate and Inter-beat Interval Computation to Diagnose Stress Using ECG Sensor Signal," 2010.

[32] S. Begum, M. S. Islam, M. U. Ahmed, and P. Funk, "K-NN Based Interpolation to Handle Artifacts for Heart Rate Variability Analysis," in *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT'11)*, 2011, pp. 387–392.

[33] 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.

[34] A.-B. Liu, P.-C. Hsu, Z.-L. Chen, and H.-T. Wu, "Measuring Pulse Wave Velocity Using ECG and Photoplethysmography," *J. Med. Syst.*, vol. 35, no. 5, pp. 771–777, 2011.

[35] D. Novak, M. Mihelj, and M. Munih, "A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing," *Interact. Comput.*, vol. 24, no. 3, pp. 154–172, May 2012.

[36] A. Haensel, P. J. Mills, R. A. Nelesen, M. G. Ziegler, and J. E. Dimsdale, "The relationship between heart rate variability and inflammatory markers in cardiovascular diseases," *Psychoneuroendocrinology*, vol. 33, no. 10, pp. 1305–1312, Nov. 2008.

[37] R. Orsila *et al.*, "Perceived Mental Stress and Reactions in Heart Rate Variability — A Pilot Study Among Employees of an Electronics Company," *Int. J. Occup. Saf. Ergon.*, vol. 14, no. 3, pp. 275–283, 2008.

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

[39] 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.

[40] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

[41] I. Kononenko, E. Šimec, and M. Robnik-Šikonja, "Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF," *Appl. Intell.*, vol. 7, no. 1, pp. 39–55, 1997.

[42] K. Li and Y. Han, "Study of Selective Ensemble Learning Method and Its Diversity Based on Decision Tree and Neural Network," in *2010 Chinese Control and Decision Conference (CCDC)*, 2010, pp. 1310–1315.

[43] N. Jaques, S. Taylor, A. Azaria, A. Ghandeharioun, A. Sano, and R. Picard, "Predicting students' happiness from physiology, phone, mobility, and behavioral data," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII'15)*, 2015, pp. 222–228.

[44] 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.

[45] X. Sun and A. May, "A Comparison of Field-Based and Lab-Based Experiments to Evaluate User Experience of Personalised Mobile Devices," *Adv. Human-Computer Interact.*, vol. 2013, pp. 1–9, 2013.

[46] R. Rawassizadeh, E. Momeni, C. Dobbins, P. Mirza-Babaei, and R. Rahnamoun, "Lesson Learned from Collecting Quantified Self Information via Mobile and Wearable Devices," *J. Sens. Actuator Networks*, vol. 4, no. 4, pp. 315–335, Nov. 2015.

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

Chelsea Dobbins received the BSc (Hons) degree in Software Engineering and the PhD degree in Computer Science from Liverpool John Moores University (LJMU) in 2010 and 2014, respectively. Currently, she is a Senior Lecturer within the Department of Computer Science at LJMU. She has been the recipient of a research grant from the UK Engineering and Physical Sciences Research Council (EPSRC) and has also received an ACM Computing Review Notable Article of 2016 award for work related to mining multivariate temporal smart mobile data. She is also currently an Academic Editor for PLOS ONE.

Stephen Fairclough received his BSc (Hons) in Psychology from the University of Central Lancashire in 1986. He obtained a PhD from Loughborough University in 2000. He is currently a Professor of Psychophysiology within the School of Natural Sciences and Psychology at Liverpool John Moores University. His 2009 paper on Physiological Computing was awarded a prize as most-cited paper in the journal Interacting with Computers in 2011 and 2012. He has co-edited four collected volumes on this topic and three special issues of journals. He is currently a member of the Executive of the Human Factors and Ergonomics Society (European Chapter).