Deep learning with wearable based heart rate variability for prediction of mental and general health

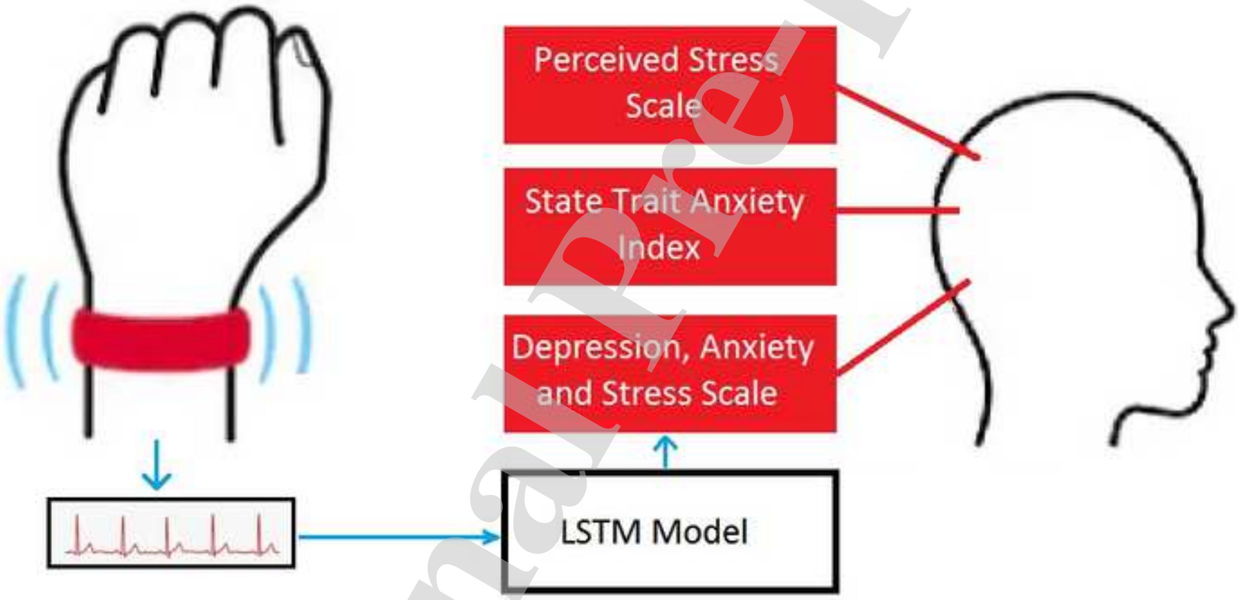
Louise V. Coutts, David Plans, Alan W. Brown, John Collomosse

PII: S1532-0464(20)30238-0
DOI: https://doi.org/10.1016/j.jbi.2020.103610
Reference: YJBIN 103610

To appear in: *Journal of Biomedical Informatics*

Received date : 29 February 2020
Revised date : 23 October 2020
Accepted date : 25 October 2020

*Graphical Abstract


Perceived Stress Scale
State Trait Anxiety Index
Depression, Anxiety and Stress Scale
LSTM Model

*Highlights (for review)

Highlights

Up to 5 points, max 85 characters each (incl. spaces)

- A novel study further exploiting data collected from wearable biosensors
- Heart rate variability (HRV) was recorded using wearables in 652 participants
- Long Short Term Memory networks link HRV to mental & general health

Deep Learning with Wearable Based Heart Rate Variability for Prediction of Mental and General Health

Louise V Coutts[1], David Plans[2], Alan W Brown[2], John Collomosse[1]

1. University of Surrey, England, 2. University of Exeter, England

Abstract

The ubiquity and commoditisation of wearable biosensors (fitness bands) has led to a deluge of personal healthcare data, but with limited analytics typically fed back to the user. The feasibility of feeding back more complex, seemingly unrelated measures to users was investigated, by assessing whether increased levels of stress, anxiety and depression (factors known to affect cardiac function) and general health measures could be accurately predicted using heart rate variability (HRV) data from wrist wearables alone. Levels of stress, anxiety, depression and general health were evaluated from subjective questionnaires completed on a weekly or twice-weekly basis by 652 participants. These scores were then converted into binary levels (either above or below a set threshold) for each health measure and used as tags to train Deep Neural Networks (LSTMs) to classify each health measure using HRV data alone. Three data input types were investigated: time domain, frequency domain and typical HRV measures. For mental health measures, classification accuracies of up to 83% and 73% were achieved, with five and two minute HRV data streams respectively, showing improved predictive capability and potential future wearable use for tracking stress and well-being.

Keywords: Machine learning, LSTM, Heart Rate Variability, Mental Health, Wearables

1. Introduction

Recent interest in inexpensive wearable technologies has led to users collecting a wealth of self quantification data. However this data is generally not made available to the individuals collecting it or is available in a limited form, generally only offering information over short time periods (e.g.

daily or hourly), for example activity levels and heart rate. In the past, 24 hour heart rate measurements were typically only recorded in a clinical setting. However with the recent pervasive use of wrist wearables, collecting and storing plethysmographic cardiac data, it is now possible to build classifying machine learning models, where long term data is available, often over months to exploit wearable technologies for real-time and long-term monitoring of health.

Whilst commonly measured variables such as activity or heart rate have been widely investigated, such a wealth of data may also be sufficient to predict seemingly non-cardiac related factors such as psychological factors affecting mental health, like stress or anxiety, that are known to affect cardiac function [7]. This study therefore aimed to assess the feasibility of using deep (machine) learning techniques on long-term data from wearables, as a method to detect issues in mental health, with a view to progressing research in this field towards providing useful user feedback and/or clinical intervention for mental health issues. If feasible, such a method could be used in an integrative manner to provide assistive technology, aiding self management and feedback (to both users and clinicians) on treatment outcomes, such as for non-pharmaceutical treatments.

For this study a Long Short-Term Memory (LSTM) network was employed, a type of deep Recurrent Neural Network designed for sequence problems such as time series analysis. As participants were undergoing exams during the time that they wore the wearable wristband, the machine learning model was developed to predict stress level, by tagging the training data with participant stress level as measured with weekly online subjective questionnaires. This initial study classified stress as either high or low, similar to the LSTM study undertaken by Umematsu et al. [20]. Participation continued after exams, providing a range of stress levels from which to train the model. The cardiac data collected was heart rate variability (or peak to peak interval of the heart beat, HRV). HRV has recently become popular as an indicator of cardiac health, showing further insight than heart rate. We hypothesised that during the exam period, stress levels would be elevated and lower HRV observed, compared to the post exam period, providing potential cues from which a machine learning model could classify stress level.

The aim of the study was to undertake a preliminary investigation of the feasibility of building machine learning models from wearable data, providing guidance for future development of similar group-wise systems and the required magnitude, precision and resolution of data required to successfully build machine learning models capable of accurately determining psychological status. The ultimate aim being to build machine learning models that can achieve sufficient accuracy to allow wearable users to be fed a regular, automatic measure of their psychological status, e.g. current stress level, without the need for any subjective input from the wearer.

2. Materials and Methods

2.1. Sample and Participants

The sample was recruited from students at the University of Surrey, England.

2.1.1. Trial 1

Participants were drawn from those undertaking undergraduate or postgraduate level exams in January 2018. One hundred individuals consented to take part, with 91 of those participants (62% female, 38% male) with an age range of 18 to 38 years (mean 21 years) completing the initial questionnaire. Due to problems with wristband data transfer, the wristband data sample size consisted of 68 participants (mean age 21 years, 64% females).

2.1.2. Trial 2

Participants were drawn from those undertaking undergraduate or postgraduate level exams in May 2018 and January 2019. 799 individuals consented to take part in trial 2, with 600 of those participants (72% female, 28% male, age range 18 to 69 years (mean 22 years) completing the initial questionnaire and 584 participants transmitting band data.

2.2. Procedure

The University of Surrey Ethics Committee gave a favourable opinion for the research and all participants provided fully informed consent before completing the initial online questionnaire. Along with general questions about medical health, mental health diagnoses, general lifestyle and satisfaction with the courses they were currently studying, the initial questionnaire assessed willingness to share data and common personality and psychological measures (results to be published elsewhere), along with: Perceived Stress Scale [6], State and Trait Anxiety [14] and Depression Anxiety Stress Scale [12], all measures previously found to correlate with exam stress [1]. All questionnaires were scored following the cited standard guidelines.

Individuals were each given a wrist band that was then paired with their phone and instructed on how to keep the app running in the background to allow regular data transfer. They were asked to wear the wristband continually (apart from bathing and charging) and behave as they would normally for the duration of the trial. No visibility over the data collected was offered to participants during the data collection period. Participants were also asked to complete a short mental health questionnaire on a weekly basis (on a weekly basis for trial 1 over a four week period and twice weekly for trial 2, over an eight week study period) which included the Perceived Stress Scale, State and Trait Anxiety and Depression Anxiety Stress Scale measures described above. Responses to each mental health survey were then converted into numerical values using the standard guidelines for each survey ([6], [14], [12]), with higher levels of stress, anxiety or depression represented with higher scores.

2.3. Heart Rate Variability Data Processing

The Biobeam band (BioBeats Group Ltd., London) was selected due to the ability to capture inter-beat intervals and from which HRV could be derived. The device was programmed to turn on every 30 minutes, capturing inter-beat intervals (IBI) for a period of 5 minutes for trial 1 and 2 minutes for trial 2, before switching off again. This meant that data could be acquired passively without user intervention and possible bias of data collection. The device sensor collected IBI data optically by pulse oximetry at a sampling frequency of 50hz. Accelerometer measurements were also captured by the device to enable IBI data points collected during high physical activity to be automatically discarded during data collection, as described in Morelli et al [16]. The band was paired through an iOS app called Biobeam (Biobeats Group Ltd., London) through the user's iPhone, which periodically collected data from the sensors and uploaded data to the data cloud for analysis.

As the daytime IBI data collected was visually different in characteristics to the data collected at night time, but no activity data was available in order to allow segmentation by activity, the data was split into two sections based on time, where ”day” data represents data collected between 8am and midnight and ”night” data refers to data collected between midnight and 8am and these two data sets were analysed separately for both trials. Similarly, where no concurrent activity data was available, it was not possible to remove activity-based noise directly from the data. However different activity levels between groups (and differences in activity-based noise) may have improved classification accuracy beyond that achievable with activity filtered IBI data, where for example data from a participant with a high score on the depression scale may have been less active than a participant with a low score.

2.3.1. Trial 1

Due to the band's heart rate sensor being affected by motion artefacts, IBI data points greater than 40% (as recommended for the participant age group by Karlsson et al, 2012 [10]) from the mean of the preceding and following IBI data points were removed from the dataset during pre-processing, by cycling through the filtering data point removal process five times. The bi-hourly 5-minute batches of data were then processed and presented to the machine learning model in 3 ways: 1. Time-domain sequences of 50 IBI interval data points (equating to a measuring period of approximately 45 seconds for daytime data and 50 seconds for night-time data), were differenced between successive data points and used to train a time based machine learning model. 2. Frequency domain sequences of IBI spectral characteristics (amplitudes of frequency harmonics between 0 and 0.5 Hz) were used to train a frequency-based machine learning model. The power of the response at different frequencies were calculated using the Matlab function "bandpower", in increments of 0.005 Hz (a sampling resolution of 100). 3. A sequence of typical HRV measures [3], [15], : (1) mean and (2) standard deviation of HRV over the 5 minute data batch, (3) 24 hour standard deviation of (1), (4) 24 hour mean of (2), (5) root mean square successive difference (RMSSD), (6) percentage of IBI intervals (within the five minute batch) greater than 50ms (pnn50), frequency power within ranges: (7) ultra (0 to 0.003 Hz), (8) very low (0.003 to 0.04 Hz) (9) low (0.04 to 0.15 Hz), (10) high (0.15 to 0.4 Hz) and (11) total (0 to 0.4 Hz) and (12) ratio of low to high frequency (L/H Ratio), also calculated using Matlab "bandpower". Data were also analysed in a standard manner to investigate differences between groups.

A maximum of 48 of these data sets were available per participant per day (twice per hour), however in practise an average of two adequately long data sets per day per participant were available for analysis. Whilst this limitation arose pre-dominantly due to: 1) problems in transmitting data through the app or 2) problems with band to phone pairing, this lack of data transmission would also have been partly due to 3) participants removing the band for showering and charging and 4) high activity during the five minute data acquisition and too few usable data points being recorded. Nevertheless, over the four-week long study period, this resulted in producing in the region of 2000 data sets for analysis, that had accompanying mental health measure labelling, derived from the results of the weekly subjective questionnaires.

2.3.2. Trial 2

Karlsson et al [10] proposed removal of IBI data points exceeding an age related limit (of between 20% and 40%), with a peak in limit of 40% at the age of 15. However Karlsson et al assessed just four age groups, without further investigating the optimum limit within the predominant participant age group for this study (15-24 years), therefore to further investigate this method, for trial 2 the limit for including IBI data points was varied between 20% and 50% at 3.75% intervals, to explore the optimal threshold limit for machine learning model accuracy. In addition, where Karlsson et al, 2012 [10] recommended cycling through the data point removal process a maximum of five times, the effect of the number of filtering cycles on machine learning classification accuracy was also investigated, by training the model using datasets which had undergone different number of filtering cycles, from 1 to 20.

The bi-hourly 2-minute batches of data were then processed and presented to the machine learning model using only frequency domain sequences of IBI spectral characteristics (amplitudes of frequency harmonics between 0 and 0.5 Hz). Where batches of data were shorter for this trial, the Lomb-Scargle technique was implemented instead of the Matlab 'bandpower' function. This technique takes account of samples collected with non-equal sampling frequency, is commonly used for HRV data analysis ([17] and [19]; calculated using the Matlab function "plomb") and is also able to produce robust outputs for smaller samples of data. In order to create data samples of equal length for machine learning processing, for the second method, the Lomb-Scargle outputs were each individually resampled to form datasets characterising frequency response between 0 and 0.5 Hz with sampling resolution 0.005 Hz).

The number of data sets collected for analysis that had accompanying mental health measure labelling derived from the results of the twice weekly subjective questionnaires for trial 2 was in the region of 500,000. This increase was due to increased participant numbers, data transmission rates and trial duration. As for trial 1, daytime and night-time data were analysed separately, where 'day' data represents data collected between 8am and midnight and 'night' data refers to data collected between midnight and 8am.

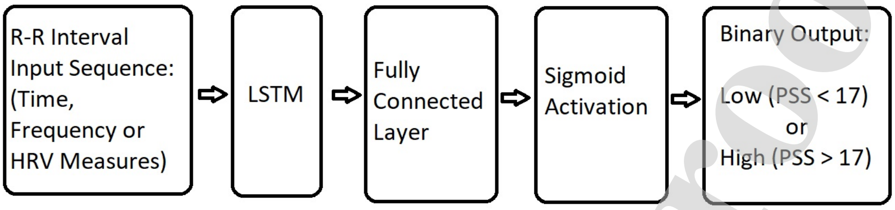


Figure 1: Network architecture for both LSTM based models (from scratch and pre-trained). Example shown for Weekly Measure: Perceived Stress Scale (PSS).

2.4. Binary Classification Using LSTM

Each data sample was tagged as either above or below a set threshold, for each of the five different measures: (1) Perceived Stress Scale (PSS), (2) State Trait Anxiety Index (STAI) and (3) Depression (DASSd), (4) Anxiety (DASSa) and (5) Stress (DASSs) Scale for both trials and also (6) sleep quality, (7) loss of appetite, (8) stomach discomfort, (9) diarrhoea, (10) suffering from cold or headache and (11) level of alcohol consumption (as reported in the twice weekly questionnaires) for trial 2. For scored scales (1-5), thresholds were set as the median score: PSS=17, STAI=15,DASSd=2.5, DASSa=3 and DASSs=5.5. For categorical scales (6 and 11), thresholds were set between positive and negative reports for sleep quality and above 8 units of alcohol per week for level of alcohol consumption. All other scales were polar (7-10), therefore tagging directly represented yes or no answers. Balanced data sets for each scale were then whitened to a mean of zero and standard deviation of 1. The datasets for each scale were then split into 3 sub-sets: training (80%), test (10%) and validation (10%), with each sub-set containing data from different groups of participants to avoid leaking data from training to test data. For all scales, each balanced subset contained data from a minimum of 2 (trial 1) or 15 (trial 2) participants.

The LSTM models were created using Keras on Python using the architecture described in Figure 1, with hyperparameters as described below. Network weights were optimised by minimising cross-entropy loss function using the 'ADAM' optimiser [11]. The LSTM hidden layer was formed of 100 dimensions, a fully connected dense ReLu layer with 132 outputs and a final output layer with sigmoid activation for binary output classification, tuned from the validation data. An LSTM model was trained for 25 epochs with a batch size of 4 for each data set (for each of the different scales).

Classification accuracy was used as a performance indicator across all of the different models investigated.

2.4.1. Trial 1

Three different models were tested, firstly where the model was trained as standard, using the data as classified by each mental health measure. Secondly, because of the relatively small data sets available, the large variation between day and night time data was used to pre-train the LSTM model. The pre-trained weights were then used as a starting point for training the model using the data as classified by each mental health measure. This second approach is a standard technique to improve classification accuracy where data is limited [4]. The third 'hot-spot' method aimed to optimise the standard model by removing local regions within training data sets that achieved low classification accuracy. This is reported for the frequency domain data only, where data for each 0.005 Hz frequency increment was first trained using the standard LSTM model and only data from frequency increments achieving high classification accuracy used for hot-spot model training.

2.4.2. Trial 2

Frequency domain data were used to train the standard version of the model as described for trial 1, using the data as classified by each mental health or subjective measure.

3. Results

3.1. Trial 1

Significant differences in HRV measures between high (PSS >median score of 17) and low (PSS <median score of 17) stress groups were observed (see table 1), especially within night time data. Significance levels between groups repeated in a similar manner across the other four subjective measures (STAI, DASSd, DASSa and DASSs). Significant differences in frequency domain data between high and low groups were also observed at low frequencies (in the region of 5 to 25 Hz) for all weekly measures except DASSa for daytime data, and for DASSs and DASSd for night time data (see Figure 2). No significant differences were found between groups in time domain data, however when median values were calculated across each one minute sample of data (across tagged data samples used for training) these median values

	Day		Night	
	PSS <17	*PSS >17*	*PSS <17*	*PSS >17*
Mean RR	0.969	0.865	1.019	0.913 *
STD RR	0.406	0.235	0.139	0.117
STD 24hr Mean	0.458	0.431 *	0.457	0.420 *
Mean 24hr STD	0.1522	0.0722 *	0.0569	0.0484 *
RMSSD	0.344	0.222	0.156	0.132 *
PNN50	23.611	24.272	23.932	23.387
Ultra Low Freq	0.905	0.731	1.041	0.825 *
Very Low Freq	0.830	0.700	1.018	0.826 *
Low Freq	0.0576	0.0160	0.0056	0.0034
High Freq	0.0336	0.0133	0.0064	0.0045
Total Power Freq	1.045	0.777	1.085	0.846 *
L/H Ratio	1.425	1.133 *	0.909	0.823 *

Table 1: Median HRV Measures for data tagged with level of Perceived Stress (PSS), classified as either above and below the threshold of 17. Units: Mean - RMSSD (seconds), PNN50 (%) and frequency (power spectral density), significant differences between PSS <17 and PSS >17 data are indicated by * (unpaired t-tests, p<0.01).

were significantly different between high and low groups for all subjective measures, for both day and night data (unpaired t-tests, p<0.001).

For machine learning binary classification, a pure random (chance) response would be 50%. For the limited data pool used for training, the time domain data was fairly poor at classifying, across all of the weekly measures. In contrast, the classification accuracy using the frequency domain data and the HRV measures (with night-time data) is higher (see Figure 3), and shows promise that further training with more data may lead to reasonable accuracy levels that have practical use in wearables. Classification accuracy for frequency domain data improved when using the pre-trained model (mean increase in accuracy: 6.9 %) for all weekly measures, except Perceived Stress Scale, whereas for time domain data and HRV measures, the pretrained model had negligible effect on classification accuracy (mean increase of: 0.0 % and -0.6 %, respectively). For all weekly measures except Perceived Stress Scale daytime data and DASSd night-time data, classification accuracy also improved when using the hot-spot method (mean increase

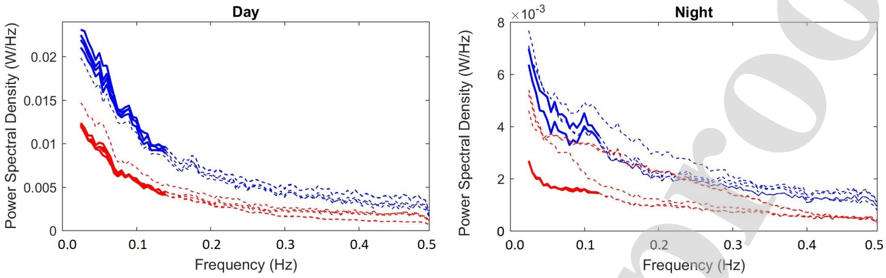


Figure 2: Median of frequency domain data for high (red/grey) and low (blue/black) weekly measure groups, shown for all five subjective weekly measures, for a) Day and b) Night time data. Bold lines indicate where power spectral density was significantly different between high and low subjective measure groups ($p<0.01$, unpaired t-tests). Significant differences between groups were observed for all subjective measures (except DASSa) in daytime data and for DASSs and DASSd in night time data.

in accuracy: 2.9%).

3.2. Trial 2

For machine learning binary classification, classification accuracy for frequency domain data across the filtering thresholds and cycles described is shown in Figures 4 and 5 for daytime and night-time data. For each of the measures recorded in the survey, optimum classification accuracies were generally achieved by undertaking a greater number of IBI filtering cycles than the maximum of five recommended by Karlsson et al, 2012 with smaller thresholds than the age related threshold of 40% also recommended by Karlsson et al, 2012, [10]. The maximum achievable classification accuracy for each measure, using the parameters tested, are described in Table 2, showing that higher classification accuracy was always achieved using Night-time compared to daytime data. Higher classification accuracies were achieved for DASSd, sleep quality, loss of appetite, stomach discomfort, diarrhoea, cold and headache and alcohol consumption, than PSS, STAI, DASSs and DASSa.

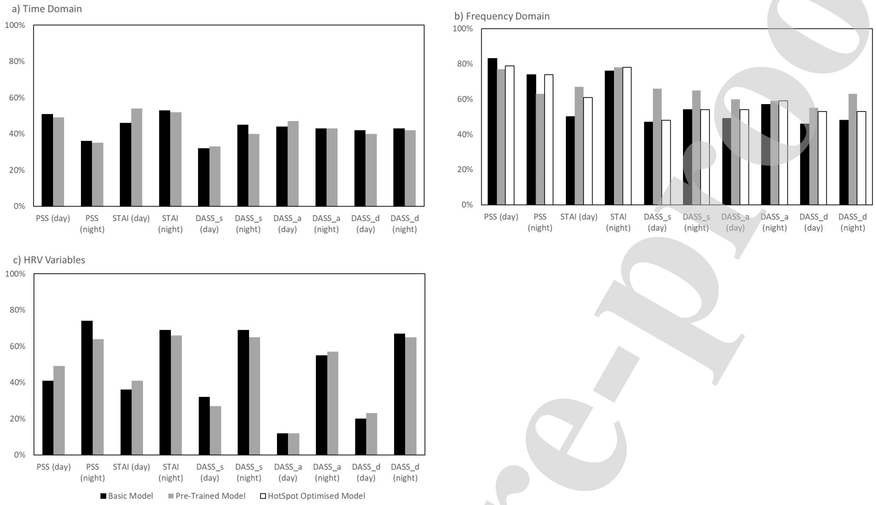


Figure 3: Classification accuracy of the from-scratch model (shown in black), pre-trained (shown in grey) and hotspot (for frequency domain only, shown in outlined white) LSTM models, for classification of the different weekly measures: Perceived Stress Scale (PSS), State trait anxiety index (STAI) and Depression, Anxiety and Stress Scale (Depression: DASSd, Anxiety: DASSa and Stress: DASSs) for the three data types: (a) time domain, (b) frequency domain and (c) HRV measures.

4. Discussion

The primary study aim was to investigate the feasibility of creating machine learning models enabling more complex feedback to be provided to wearable users. The example case investigated used HRV data collected from wrist wearables to predict stress, anxiety and depression and general health. Classification accuracy levels of up to 83% for Trial 1 (see Figure 3) and 85% for Trial 2 (see Table 2) were achieved, sampling data from just 5 minute and 2 minute data windows respectively and is comparable to that achieved by Umematsu et al, who used data collected over much longer periods, up to 7 days [20]. The maximum classification accuracy for mental health scores (PSS, STAI and DASS) was 83% for trial 1 and 73% for trial 2. This difference is likely due to the reduction from five to two minute data stream length between the first and second trials.

	Day	Night
PSS	61.7%	63.8%
STAI	63.4%	63.9%
DASSs	64.3%	67.7%
DASSa	63.2%	69.4%
DASSd	71.5%	72.6%
Sleep Quality	70.3%	74.5%
Loss of Appetite	82.4%	85.0%
Stomach Discomfort	64.7%	77.5%
Diarrhoea	73.7%	75.0%
Cold or Headache	67.5%	72.2%
Alcohol Consumption	71.6%	75.9%

Table 2: Maximum classification accuracy for each of the subjective survey measures recorded, for the range of IBI filtering parameters tested.

For trial 1, higher classification accuracy was achieved for PSS and STAI than DASS measures, using 5-minute duration data streams, possibly reflecting the reduction in significant differences found between groups within the frequency domain data for DASS measures (see Figure 2). For trial 2, where 2-minute data streams were collected, the opposite was true, with higher accuracy achieved for DASS measures than PSS and STAI. This may reflect differences in the effects of stress, anxiety and depression on cardiac behaviour and also highlight that whilst different survey measures may report measuring the same feature, e.g. where both DASSs and PSS measure stress, they involve different survey questions and consequently measure different aspects of the selected feature.

When measured statistically, the significant differences in HRV measures with mental health measures (see figure 2) also provide extra evidence that the typical measures used to analyse HRV are diagnostically valuable. Given the findings from trial 1 that the classification accuracy using frequency domain data and HRV measure data on average outperform the classification accuracy using time domain data, this suggests that further processing from the time domain, before inputting data into the machine learning model, is worthwhile.

Whilst classification accuracy was higher for all techniques investigated in this study when training models on the Night-time data compared to the daytime data, the different measures investigated did not display such repetitive trends. This indicates that the predictive capability for each individual measure should be optimised by separately appraising a range of machine learning models, including those used in this study. Both the pre-trained and hot-spot methods improved accuracy in some data sets, therefore are techniques worth considering, but should be compared against a model trained from scratch, to ensure optimum accuracy is achieved for each scenario (data type, subjective measure, etc). Combining these two methods and further optimisation of LSTM model architecture may further improve accuracy. For example, using a triplet (e.g. siamese) architecture to explicitly learn from 'hard negative' examples that are challenging to discriminate [13].

Further, the findings from trial 2 demonstrate that the filtering process is an important feature in determining model accuracy and should be evaluated for each individual measure being processed. The general finding that maximum classification accuracy can be achieved by using smaller filtering thresholds and greater numbers of filtering cycles than previously reported by Karlsson et al [10] may reflect differences in the accuracy of the measurement device, where wearable devices likely achieve lower accuracy than devices used for clinical studies. Further development of machine learning models, especially those typically classified as 'deep learning' models, may be sufficient to replace the pre-processing used in the study, where a deep learning model would learn to compute and use similar filtered HRV and frequency measures for classification within the model itself. For example, making use of the significant differences between groups observed by statistical analysis in this study, at low frequencies (see Figure 2) and in HRV measures (see Table 1).

Participants with a higher body mass index (BMI) would be more likely to suffer with conditions related to cardiovascular disease [5] and consequently wearable data collected from this group may differ from those with lower BMI. The impact of these differences (and other possible co-factors such as gender) on the classification accuracy achieved from machine learning models was not investigated as part of this study, where separate machine learning models trained with data from groups of participants of similar BMI may result in improved classification accuracy. In addition, whilst the optimisation investigation conducted within this study employed set filtering parameters for the entire data set, subject specific filtering parameters have been recommended based upon the age of the subject [10], therefore classification may be further optimised by employing a combined approach to the filtering process.

Nevertheless, the study has shown that with reasonable accuracy, the use of deep learning models predicting different aspects of general and mental health is feasible using HRV data from wrist wearables. In future, such models may be built to be explainable, enabling the methods developed in the model to be recorded and reported back, meaning that such models may enable new features to be found within HRV data that may aid clinical diagnosis.

5. Conclusion

The study investigated the feasibility of using HRV data collected from wrist wearables to predict both general health and mental health measures and demonstrated that reasonable accuracy could be achieved with relatively small training data volumes using contemporary deep learning techniques (LSTM). This is particularly the case when the LSTM is (1) pre-trained on a simpler classification task (e.g. day night discrimination) for which greater data volumes are available, and subsequently fine-tuned for the target survey measure e.g. stress, and (2) tuned with optimal filtering parameters. Classification accuracy of up to 85% was achieved using just two- or five-minute data streams which was comparable to previous studies, where LSTMs were trained using data collected over days rather than minutes. The health measures investigated in this study are just a few of the possible measures worth investigating and demonstrate huge potential for providing more complex feedback to wearable users and companies issuing wearables for health tracking, than is currently provided in everyday commercial products.

Acknowledgment

This research has been funded by the UK Engineering and Physical Sciences Research Council grant: Co-operative Models for Evidence-based Healthcare Redistribution (EP/P03196X/1).

Data Availability

The data collected in this study resides in a secure network and access to data for further analysis would require further ethics approval due to the data containing sensitive participant information, but may be available upon request.

References

[1] E. Austin and D. Saklofske and S. Mastoras, Emotional intelligence, coping and exam-related stress in Canadian undergraduate students, Australian Journal of Psychology: 62(1), 42-50,2010.

[2] R. Bar-On, Bar-On Emotional Quotient Inventory: Short. Technical manual, Toronto: Multi-Health Systems, 2002.

[3] R. Castaldoa, P. Melillob, U. Bracalec, C.M. Casertaa, M. Triassic and L. Pecchiaa, Acute stress assessment via short term HRV analysis in healthy adults; A systematic review with meta-analysis, Biomedical Signal Processing and Control, 2015.

[4] Choi E, Schuetz A,Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset J Am Med Inform Assoc, 2017, 24(2): 361–370. doi: 10.1093/jamia/ocw112

[5] M. Chughtai, A. Khlopas, JM Newman, GL Curtis, N Sodhi, PN Ramkumar, R Khan, S Shaffiy, A Nadhim, A Bhave, MA Mont. What is the Impact of Body Mass Index on Cardiovascular and Musculoskeletal Health? 1em plus 0.5em minus 0.4emSurg Technol Int. 2017 Jul 25;30:379-392.

[6] S. Cohen, T. Kamarck and R. Mermelstein, A global measure of perceived stress Journal of Health Social Behavior, 24, 385-396, 1983.

[7] R.K. Dishman, U.Y. Nakamur, M.E. Garcia, R.W. Thompson, A.L. Dunn and S.N. Blair, Heart rate variability, trait anxiety, and perceived stress among physically fit men and women, International Journal of Psychophysiology,37, 121-133, 2000.

[8] N.S. Endler and J.D.A. Parker, Coping Inventory for Stressful Situations (CISS): Manual, Toronto: Multi-Health Systems, 1999.

[9] M. Franceschi, D. Morelli, D. Plans, A. Brown, J. Collomosse, L. Coutts and L. Ricci. ComeHere: exploiting Ethereum for secure sharing of health-care data. Workshop Paper, 24th International European Conference on Parallel and Distributed Computing, Turin, August 2018.

[10] M. Karlsson, R. Hörnsten, A. Rydberg and U. Wiklund, Automatic filtering of outliers in RR intervals before analysis of heart rate variability in Holter recordings: a comparison with carefully edited data, BioMedical Engineering OnLine, 11(2), 2012.

[11] D.P. Kingma and J. Ba, Adam: a method for stochastic optimisation, http://arxiv.org/abs/1412.6980, Arxiv, 2014.

[12] S.H. Lovibond and P.F. Lovibond, Manual for the Depression Anxiety Stress Scales, 2nd Ed. Sydney: Psychology Foundation, 1995.

[13] F. Radenović, G. Tolias and O. Chumřej, CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples, Proc. ECCV, 3-20, 2016.

[14] T.M. Marteau and H. Bekker, The development of s six-item short form of the state scale of the Sielberger State-Trait Anxiety Inventory (STAI), British Journal of Clinical Psychology, 31, 301-306, 1992.

[15] J.E. Mietus and A.L. Goldberger, Heart Rate Variability Analysis with the HRV Toolkit https://physionet.org/tutorials/hrv-toolkit/, 2014.

[16] D Morelli, L Bartoloni, M Colombo, D Plans, DA Clifton, Profiling the propagation of error from PPG to HRV features in a wearable physiological-monitoring device, Healthc Technol Lett. 2018, 5(2):59-64.DOI: 10.1049/htl.2017.0039

[17] Piskorski J, Guzik P, Krauze T, Żurek S. Cardiopulmonary resonance at 0.1 Hz demonstrated by averaged Lomb-Scargle periodogram, Central European Journal of Physics, 2010, 8(3):386-392. DOI: 10.2478/s11534-009-0101-1

[18] G. Saucier, Mini-markers. A brief version of Goldberg unipolar Big Five markers, Journal of Personality Assessment, 1994, 63, 506-516, 1994.

[19] Simões Fonseca D, Netto A, Bartels Ferreira R, Maurício Ferreira A, Miranda de Sá L. Lomb-scargle periodogram applied to heart rate variability study Biosignals and Biorobotics Conference (BRC), 2013. DOI: 10.1109/BRC.2013.6487524

[20] T. Umematsu, A. Sano, S. Taylor and R. Picard, Improving Stress Forecasting using LSTM Neural Networks, The 40th International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2018.

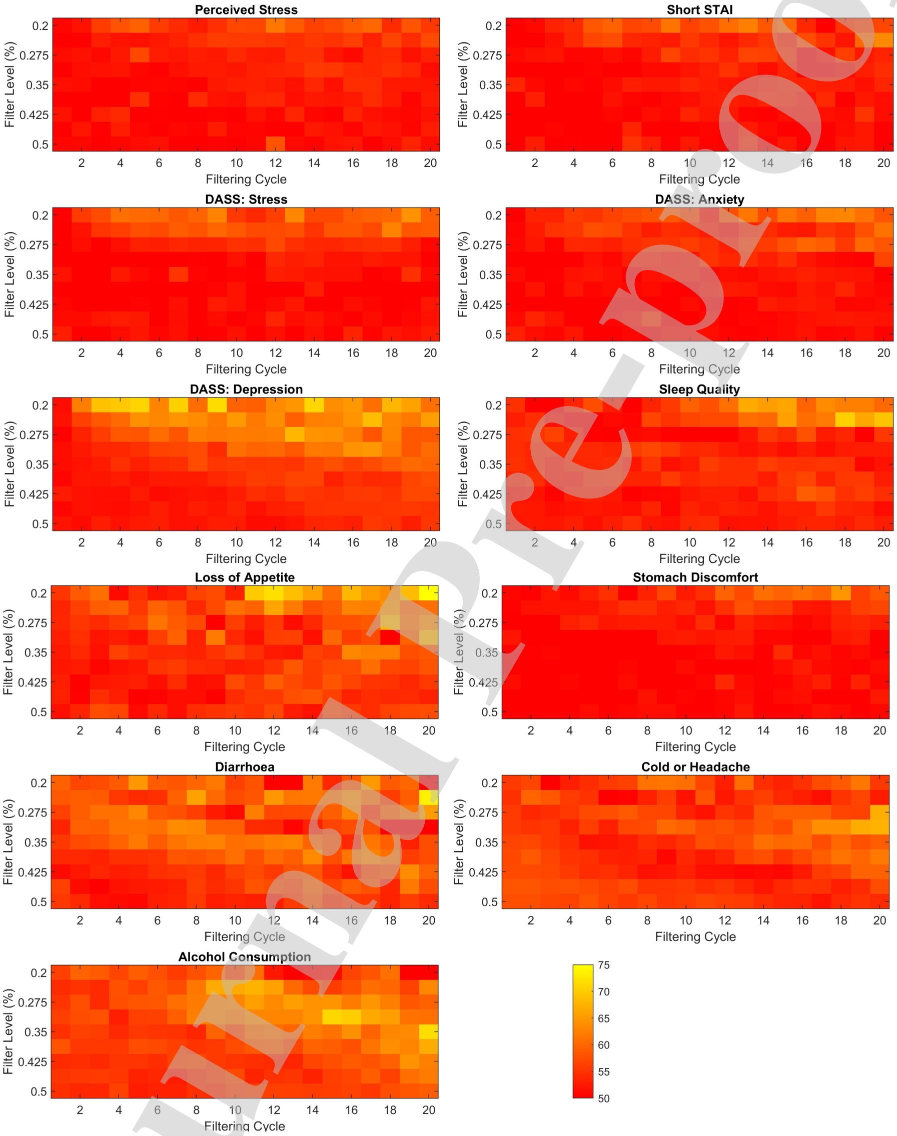


Figure 4: Classification accuracy of the from-scratch LSTM models, for classification of the different measures recorded in the subjective survey, using frequency domain Daytime IBI data where IBI data points exceeding thresholds of between 0.2% and 0.5% were pre-filtered between 1 and 20 times. Classification accuracy represented by greyscale level (lower accuracy shown in black, higher accuracy in white).

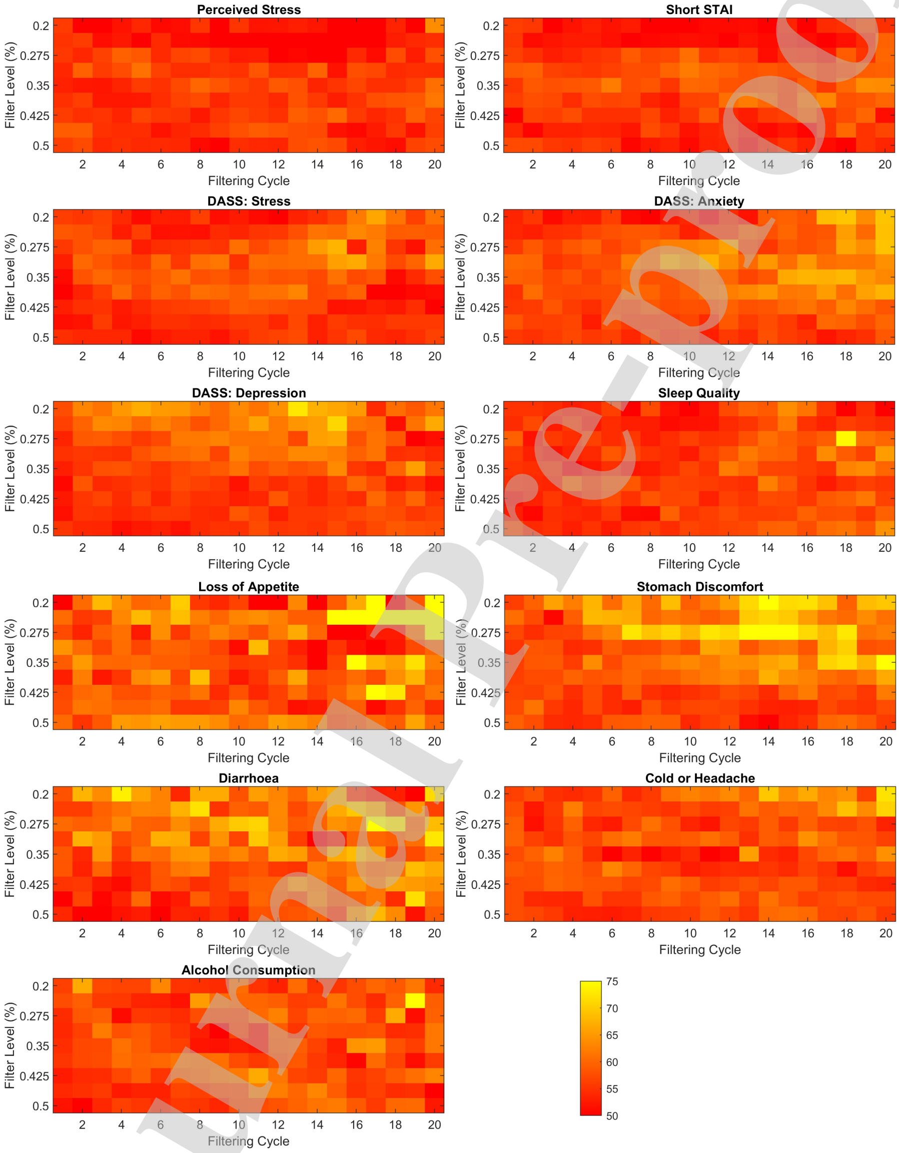


Figure 5: Classification accuracy of the from-scratch LSTM models, for classification of the different measures recorded in the subjective survey, using frequency domain Night-time IBI data where IBI data points exceeding thresholds of between 0.2% and 0.5% were pre-filtered between 1 and 20 times. Classification accuracy represented by greyscale level (lower accuracy shown in black, higher accuracy in white).

CRediT author statement

Louise Coutts: Conceptualization, Methodology, Data Curation, Writing - Original Draft Preparation, Writing - Review & Editing, Visualization, Software, Validation, Formal analysis, Investigation

David Plans: Conceptualization, Methodology, Funding acquisition

Alan Brown: Supervision, Conceptualization, Methodology, Funding acquisition

John Collomosse: Supervision, Conceptualization, Methodology, Funding acquisition

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: