Development of a Real-Time Stress Detection System for Older Adults with Heart Rate Data

Patrizia Di Campli San Vito
Patrizia.DiCampliSanVito@glasgow.ac.uk
University of Glasgow
Glasgow, Scotland, United Kingdom

Gözel Shakeri
Goezel.Shakeri@uni-oldenburg.de
University of Oldenburg
Oldenburg, Germany

James Ross
2251463R@student.gla.ac.uk
University of Glasgow
Glasgow, Scotland, United Kingdom

Xiaochen Yang
Xiaochen.Yang@glasgow.ac.uk
University of Glasgow
Glasgow, Scotland, United Kingdom

Stephen Brewster
Stephen.Brewster@glasgow.ac.uk
University of Glasgow
Glasgow, Scotland, United Kingdom

ABSTRACT

Stress is one of the factors considerably contributing to older adult's decreasing overall health. Detecting stress in real-time could aid family members to intervene more timely and keep older adults healthier. However, many stress detection systems are not detecting in real-time, depend on multiple devices, capture a plethora of inconveniently sampled data, or use data from younger adults. In this paper, we built a real-time stress detection system for older adults using only heart beats per minute (BPM), which can be easily obtained with most single, comfortable devices. We collected data from people over 60 (N=15), evaluating the Mannheim Multicomponent Stress Test (MMST) for older adults, then built a machine learning model with a classification performance of 76% (AUC) on BPM alone and tested it in real-time in another experiment, comparing the model's effectiveness with four different heart rate devices. Detection performance decreased considerably (51%) when using the model in another experiment and could not be used successfully with other devices, while a reduced MMST induced stress comparable to the full test suite.

CCS CONCEPTS

• **Human-centered computing** → *User studies*; • **Applied computing** → **Health care information systems**.

KEYWORDS

stress detection; physiological data; machine learning

ACM Reference Format:
Patrizia Di Campli San Vito, Gözel Shakeri, James Ross, Xiaochen Yang, and Stephen Brewster. 2023. Development of a Real-Time Stress Detection System for Older Adults with Heart Rate Data. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '23), July 5–7, 2023, Corfu, Greece.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3594806.3594817

PETRA '23, July 5–7, 2023, Corfu, Greece

1 INTRODUCTION

The proportion of adults over 60 within the world's population is rising globally and is expected to reach 22% by 2050, an increase from 12% in 2015[1]. A majority of older people prefer *ageing in place*, a term used to describe the ability to grow old within the community and with more independence than is generally possible in a residential care home [38]. Further, with an increasing older population, care homes would need a significant increase in government funding to provide enough places and support[2]. The prevalence of health conditions increases with older age [28], which can in turn increase the stress level of older adults [11]. This risk is heightened for elderly in situations with low social support [4] such as when they live alone [8], which one in three older adults do [13]. Stress has been shown to increase cognitive decline [25], starting a cycle of decreasing health. If these stressful situations could be identified in real-time, a family member or carer could intervene, breaking this cycle and helping to keep the older adult more healthy, potentially increase their life span.

Technology could significantly improve stress detection and aid carers to intervene earlier, and stress detection systems have been researched in the past [11, 20, 26, 39]. However, stress detection in previous research often depended on collecting several physiological signals to detect stress, such as electromyography, galvanic skin response (electrodermal activity) [11, 26], or respiration [20]. Relying on these many different input modalities often increases the accuracy of the model, but often needs a complicated or bulky setup to collect or weeks of training data for a specific individual [36]. Especially older adults are often reluctant to use technology that inconveniences them [27] and such an extensive setup could lead to low adaption in this population group. Heart rate data alone is more easily available and many different kinds of devices can collect it in some form, with research so far mostly relying on heart rate variability (HRV) data [5, 7, 9, 34]. However, often beats per minute (BPM) are the only readily available data, especially on devices which would be convenient enough for long time use and with options for real time access, as devices capable of collecting HRV data can usually not be accessed by developers outside of proprietary software and/or not in real-time, which is both essential

[1] https://www.who.int/news-room/fact-sheets/detail/ageing-and-health (accessed 27/01/2023)
[2] https://www.gov.uk/government/publications/care-homes-market-study-summary-of-final-report/care-homes-market-study-summary-of-final-report (Report from 2017, accessed 27/01/2023)

for the presented research in this paper. For successful long-term use in the population of older adults, it is also essential that the devices are easy to use and can be removed easily without assistance, such as wrist-worn or in-ear devices. Therefore, a trade-off between a less accurate machine learning model, but easy available and convenient hardware might have to be struck for older adults. Additionally, these systems often offer no real-time stress detection and build the machine learning models on data collected from a different population group, such as young people [26] or older adults with cognitive impairments [11]. These population groups could show differences in how they react physiologically to stress and a model trained on healthy older adults could work better for them.

We have conducted two studies to build a real-time stress detection tool, using only easily available heart rate beats per minute (BPM) data and calibrated on older adults. The first study collected labelled heart rate data of older adults during induced stress, on which a machine learning model was trained to detect stress in real-time. The Mannheim Multicomponent Stress Test (MMST) [22] was used to induce stress safely, exploring its usefulness for older adults. A second experiment then tested the trained machine learning model in real-time, while stress was induced with an adapted MMST. In this experiment, heart rate was collected with four devices to compare the effect of a heart rate monitor on the data and the performance of the model.

We found that the adapted MMST was sufficient to induce stress in older adults, though in both experiments some participants were not triggered by the stressors chosen. The machine learning model built on heart rate data could detect stress with 76% accuracy in terms of AUC (area under curve) after the first experiment, but dropped significantly in the second to 51%. This is likely due to the variation in collected heart rate data in the follow-up study, which differed significantly from the first study. The comparison of the devices also showed differences in reliability of the heart rate collection. The heart rate sensors used in this research do not seem to collect similar enough data to make sure that one model could be used by several devices.
The results can help inform future research on agitation detection to create intervention measures for older adults and highlight possible pitfalls when building a real-time stress detection system.

2 RELATED WORK

Stress detection systems are built on physiological data (such as heart rate, electrodermal activity or brain activity) or behavioural data (such as speech, task evaluation or questionnaires) [6].
Physiological data for stress detection has to be collected during clearly labelled stressed and non-stressed time frames and inducing stress can be achieved with physical or psychological stressors. Physical stressors involve external environmental or internal physiological conditions of the body. Typically hot or cold temperatures or noise are used to induce stress physically. Psychological stressors can involve cognitive stressors, such as time and work pressure or isolation, which affect thought processes, or emotional stressors, such as fear- and anxiety-inducing threats or financial problems, which affect feeling responses [3].

Heart rate (HR) is often used to determine stress, typically through electrocardiograms (ECG), which can provide information on heart rate variability (HRV) and interbeat intervals (IBI) in addition to beats per minute (BPM), and can be used in the time- and the frequency- domain [5, 7, 9, 34]. HR data is often combined with additional data, such as electrodermal activity (EDA), also called galvanic skin response (GSR), skin temperature and/or respiration, to increase accuracy of the model [2, 20, 29, 37]. These studies usually involve the use of several devices to collect the data. We will focus our literature discussion on work that either relies on heart rate data alone, investigates short- or real-time aspects of stress detection or include device comparisons.

Boonnithi and Phongsuphap [5] investigated several heart rate variability features in time and frequency domain to model stress and found that mean heart rate in combination with other features such as normalised low frequency and the differences of normalised low and high frequency were useful to identify stress. The authors presented data of six subjects, but they never gave any information on them or how the stress had been induced, measured or labelled.
Investigating several features of HRV, both in time and frequency domain, in addition to morphological variability of the ECG signal, Costin et al. [9] confirmed the features identified by Boonnithi and Phongsuphap as effective for stress detection. They could increase the accuracy of the stress detection model from 80% to 90% by combining HRV features with morphological variability features. They used an online database of heart rate data collected of (not further specified) drivers in several driving situations with different stress levels due to different traffic conditions.
Hovsepian et al. [20] built a system called cStress, using heart rate data and respiration to detect stress in real-time (every 5s). Their setup included a breast strap and Android phone, and the stress was induced in the lab using cognitive, physical, and social stressors. They captured and evaluated data in a lab setting and then employed the system in the wild, validated by self-reported stress values. The authors used several lab studies to collect training and testing data individually, but did not discuss any specifics on the participant pool, such as age. Their model accuracy was 89% during lab testing and went down to 72% in the wild. They found that the model detected stress with a lower accuracy in the training phase when using heart rate data alone (78%), compared to also using respiration data (96%).
Castaldo et al. [7] investigated the usefulness of HRV analysis for a ultra short time frames (2min) to detect stress. They used the Stroop Colour Word Test (SWICT) and a fast paced video game on young adults (20s) to induce stress for development of an automated stress detection system, which was built on real-life stress events. Their model reached an accuracy of over 60%, less than previous research, and the authors argue that this discrepancy could be either due to the shorter time frame (previous research typically looked at 5min of heart rate) or the lab-based stress induction.
However, Egilmez et al. [14] found lab-studies effective for the collection of labelled heart rate data. They compared a breast-worn heart rate monitor with a wrist-worn device for stress detection of college students. The authors compared an Android smartwatch, a customised smartwatch with EDA sensor, a finger-based commercial EDA sensor and a chest-based heart rate monitor. The heart rate data was collected from 9 participants in a lab setting with several stress-inducing activities, including the ice bucket, the SWICT, a math test, a scary game and singing as psychological stressors. They

found that chest- and wrist-worn devices could detect stress equally effectively and their minute-based classification model (88.8%) outperformed their event-based model (78.8%). Singing was found to be the best stress-inducing activity in their experiment. They did not test their model in a user study after building it.

Research in this area very often focuses on building the machine learning model for stress detection, not giving much information on how and from whom the heart rate data was collected and labelled. If participant data is included, it often shows that the heart rate data was collected from young people. They also mostly do not involve a real-time application or practical evaluation of the model in another user study.

3 EXPERIMENT 1

This first experiment was conducted to test the usability of the Mannheim Multicomponent Stress Test (MMST) to induce stress in the specific population group of older adults and to collect labelled heart rate data to build the machine learning model (described in Section 4). The heart rate data was collected with an in-ear heart rate sensor, which was chosen as the aesthetics are comparable to hearing aides, which older adults are often familiar with, and as they can be removed and applied easily wothout assistance.

3.1 Experimental Setup

The Mannheim Multicomponent Stress Test (MMST) is a standardised and extensively used tool to induce stress safely [3, 22]. Heart rate was recorded 5min before and while participants executed the MMST. The MMST consisted of three variable stressors and two constant stressors, which will be described in detail. As the MMST offers these different ways of stress induction, we obtained more varied data points on stress levels.

Between the conditions, participants rated their current, subjective stress level on a 100-point Likert scale (0 = not stressed at all, 100 = very stressed) and subsequently rested for 5min. Participants were debriefed on all the deceptions at the end of the study. The study lasted approximately 1h and participants were compensated with €10, independent of how they performed during the tasks.

Variable Stressor 1: The Paced Auditory Serial Addition Task (PASAT). The computerised version PASAT-C [12] presented a number between 0 and 20 in the middle of the screen. On the left and right of the screen, rows of 0-10 and 11-20 were depicted, respectively, see Figure 1(a). The participants had to add the last two displayed numbers together and click the corresponding value on the side of the screen. This task induces a high working memory load as the sum of the number is supposed to be "forgotten" instantly, and the last shown number needs to be added to the next number. After a 5min test phase, the first 2min of the full test depicted numbers with a latency of 3s, the next 2min with a latency of 2s, and the last minute with a latency of 1s. The order of the next number was random. Shortening the display duration twice created a sensation of lack of control.

Variable Stressor 2: Stroop Colour Word Interference Test (SCWIT). During the SCWIT [32], a word was displayed in the middle of the screen, where the colour of the word randomly matched or mismatched the meaning of the word, e.g. green, see Figure 1(b). Initially, the participant had a time window of 1.9s to classify the

(a) PASAT test interface. (b) SWICT test interface.

Figure 1: PASAT and SWICT test interface.

colour of the word by saying it out loud, before the next combination was presented. Then, the presentation time was shortened up to 0.9s at the end of the test after 5min. A total of 300 colour-word combinations were presented in each condition. Similar to the PASAT-C test, a sensation of lack of control was created.

Variable Stressor 3: Affective Images (AffIm). We used images from the Geneva Affective PicturE Database (GAPED) [10], the DIsgust-RelaTed-Images (DIRTI) database [18], and Military Affective Pictures System (MAPS) [17]) as visual emotional stimuli. Four specific negative contents were included from these databases: spiders, snakes, and scenes that induce emotions related to the violation of moral and legal norms (human rights violation or animal mistreatment). Positive pictures were also included, mainly human and animal babies and nature scenes. Neutral pictures were excluded, as positive pictures provided a better contrast to the negative pictures. After a block of five negative images, one positive image was presented to avoid habituation. The negative images were presented for 5s and the positive pictures for 3s (10 positive and 50 negative pictures). None of the images was repeated, however, to avoid lack of concentration, the participants were told otherwise and were instructed to verbally say if a picture was presented twice.

Constant Stressor 1: Acoustic Stressor: white noise[3]. White noise was presented to the participants via headphones with slowly increasing volume (78dB(A) −93dB(A)). The increase in volume shall avoid any habituation during the stress induction phases and the noise would induce stress physically.

Constant Stressor 2: Motivational Stressor: loss of money. The participants were informed that the amount of money they will receive at the end of the study would depend on the number of correctly executed cognitive tasks. They were told they will receive €10 for 100% of correctly executed tasks, inducing stress emotionally. The participants received the full amount of €10 at the end of the experiment, independent of the number of correct answers.

3.2 Apparatus

The lab was equipped with a standard desktop machine on which the MMST ran. The MMST was displayed on a 28-inch Dell monitor. All participants sat on a padded chair, adjusted to their comfort. Throughout the study, participants wore a Cosinuss Two in-ear sensor [4] (see Figure 2(b) in Experiment 2, second from left), which recorded their heart rate (HR) [21] and was connected to the computer via Bluetooth. The heart rate data collected via in-ear sensor has been shown to be comparable to ECG heart rate data [1] and was logged on the computer.

[3]Pure Noise 1 from https//mc2method.org/white-noise (accessed 28/01/2020)
[4]https://store.cosinuss.com/products/cosinuss-two?variant=32175832924242 (accessed 19/07/2022)

3.3 Participants

15 healthy older adults over 60 (mean: 63.2, STDV: 2,83; 7 females, 8 males) were recruited for this study through flyers in an optometrist's practice. Participants were excluded from participation in the study if they reported having visual impairments (not including wearing glasses or contact lenses), auditory impairments, personality disorder, suffering from post-traumatic stress disorder, brain trauma, etc; heart disease or epilepsy; neurological, neuroendocrinological, or dermatological problems, consumption of caffeine beverages 2 hours prior to experiment or engagement in intensive physical exercise 2 hours prior to experiment.

3.4 Experimental Protocol

When participants first arrived, they received an information sheet explaining the experiment. After giving their consent, the heart rate sensor was attached and participants were asked to relax for 5min. This acclimatisation phase accustomised the participants to the device and alleviated any stress that participants might have been experiencing due to the setup. Afterwards, participants underwent a screening via the State-Trait Anxiety Inventory Questionnaire (STAIQ) [31]. If they scored higher than the cut-off threshold for older adults of 54/55 [24], they were excluded from participating in the study, as further inducing stress in already stressed adults was deemed unsafe. They then completed a demographics questionnaire capturing age and gender. Finally, participants were asked to rate their current, subjective stress level on a 100-point Likert scale (0 = not stressed at all, 100 = very stressed). Participants underwent all three conditions in counterbalanced order, using a Latin square. Each condition consisted of both *constant* stressors and one *variable* stressor. Our college's Ethics Committee approved the study design.

3.5 Results

The aim of this experiment was to test if stress was induced by the MMST, which we captured with the self-reported stress ratings, and to collect heart rate data for the machine learning model. We will discuss the stress ratings, descriptive heart rate data and study observations in this section and the data analysis and preparation for the machine learning model will be described in Section 4. We will not discuss participant's errors during the stress tests itself.

3.5.1 Stress Ratings. Participants rated their stress level before and after each stress test on a Likert scale with a single value between 0 and 100. The differences between the pre- and post-test ratings can be seen in Table 1. The two pre-test ratings for participant P01 for the SWICT and the AffIm were unfortunately not logged properly. As can be seen in Table 1, the Affective Image (mean: 18.53, STDV: 10.34) test only once led to a stress rating over 14, several values were even negative, showing less stress after the test than before. The math and the colour test both resulted in stressed and non-stressed states, with the PASAT (mean: 28.29, STDV: 27.34) being more effective in producing stress events than the SWICT (mean: 24.54, STDV: 18.93), both in terms of how many stress events were logged and how high the stress was rated. We defined a stress event as being above 25, corresponding to the rounded mean value of the SWICT (as the lower one between the two more effective tests) and 16 stress events were logged in this experiment.

PartID	PASAT		SWICT		AFFIM	
	Stress	HR (STDV)	Stress	HR (STDV)	Stress	HR (STDV)
P01	**68**	**69.70 (6.50)**	N/A	66.92 (7.28)	N/A	65.22 (5.30)
P02	1	68.79 (2.22)	0	66.95 (2.31)	0	N/A
P03	**45**	**75.25 (6.42)**	5	75.75 (8.19)	-5	67.23 (12.09)
P04	**32**	**67.99 (5.85)**	**56**	**65.54 (6.69)**	0	65.30 (8.13)
P05	**62**	**63.17 (3.59)**	**43**	**60.84 (5.32)**	13	61.25 (3.89)
P06	13	85.78 (14.75)	**37**	**97.14 (6.16)**	-3	89.92 (6.66)
P07	**36**	**68.79 (6.61)**	**28**	**64.40 (8.22)**	0	70.02 (4.03)
P08	**28**	**65.13 (3.80)**	6	67.17 (3.37)	2	61.81 (3.41)
P09	16	71.84 (5.89)	**31**	**73.23 (6.62)**	-12	73.18 (4.02)
P10	**26**	**57.43 (6.24)**	4	63.08 (8.76)	14	66.13 (8.73)
P11	5	75.55 (3.77)	**26**	**74.84 (6.84)**	-10	72.43 (3.06)
P12	**77**	**77.94 (6.54)**	**53**	**74.19 (4.70)**	**26**	**70.33 (3.88)**
P13	3	N/A	18	71.24 (3.71)	3	69.42 (3.10)
P14	-16	77.80 (5.64)	12	68.93 (8.11)	-2	71.64 (7.72)

Table 1: Stress Rating and Heart Rate Data of Exp. 1: showing the differences between the Likert rating (0-100) before and after each test; differences of more than 25 were defined as stress event (used for machine learning model) and the mean heart rate and standard deviation for each stress test; stress events are being shown in bold.

3.5.2 Descriptive Heart Rate Data. An overview of the mean and standard deviation of the heart rate for each participant can be seen in Table 1, where the heart rate during tests with a stress rating difference higher than 25 (see Table 1) are shown in bold font. The heart rate was not properly logged during two tests, the AffIm of P02 and PASAT of P13. Mean and standard deviation during stressful tests do not seem to follow any obvious pattern when compared to non-stressful test events.

3.5.3 Observations. During the experiment, it could be observed that participants reacted very differently to the different tests. This difference was later commented on by some participants. The affective picture task only led to one stress event, on which the participant later commented that it was triggered by a picture reminding them of a traumatic childhood event. Reactions to the PASAT of some participants depended on their enjoyment of mathematics in general. Participants who enjoyed mathematics would get invested in the results and upset when they made a mistake, while others would not get invested, rather take a break after a mistake was made and try again when they felt ready.
Regarding the constant stressors: participants commented on how they did not care about the monetary compensation, several were even reluctant to take it at all.
The collection of heart rate data with the Cosinuss Two device relied heavily on the correct positioning within the ear and was unstable in two cases, resulting in the data loss for two stress tests.

3.6 Discussion

The Mannheim Multicomponent Stress Test was used for this experiment with older adults, using the PASAT, a mathematical test, the SCWIT, a colour word interference test, and affective pictures as the variable stressors. The tests were not equally effective in triggering stress events in our participants. The affective pictures only led to a single stress event, suggesting that this test was not suitable for stress inducement in older adults. The effectiveness of

the PASAT and SCWIT varied between individuals, but in most cases at least one of the two tests would lead to a stress event. Participants reported that they were not interested in receiving money, rendering the monetary incentive as constant stressor ineffective. We, therefore, suggest that the use of the PASAT and the SCWIT in combination with the increasing white noise should be enough to induce stress in older adults. As the Cosinuss Two was not always reliable in its data collection, we plan to test several heart rate monitors in the future.

4 DATA ANALYSIS FOR MACHINE LEARNING MODEL AFTER STUDY 1

In this section, we will discuss the procedure of data processing and generation of target variable and features, then explain methods used for reducing feature dimension, handling class imbalance and performing classification, and finally present and discuss the results.

4.1 Data Processing and Feature Extraction

4.1.1 Data Preparation. Experiment 1 collected data from 14 participants under three variable stressors, leading to 42 samples. However, due to the missing of self-rated stress level (2) and/or physiological data (2), only 38 samples could be used for analysis. Each sample included HR information over a period of 5min. Considering the goal of this study was to detect stress in real time, we used a shorter duration of 1min by selecting data from 30s to 90s before the end of the experiment. The choice was made since according to the peak-end rule, the way people judge an experience are heavily weighted by how they felt at the peak moments and at the end [15]. In addition, it avoided using unrepresentative data arising from situations such as the experimenter recorded the end of experiment later than the actual end, or the participant sensed the end of the experiment based on the action of the experimenter.

4.1.2 Stress Data Processing and Stress Definition. The target variable was created by converting the continuous stress level into binary classes, stressed (S) or non-stressed (NS). Recall that a higher value on the Likert scale indicated that a subject was more stressed. Accordingly, we defined the class based on the difference in self-reported stress level before and after the test, i.e., stress (after) − stress (before). If the difference was larger than a threshold, the sample was classified as stressed (see Table 1 bold); otherwise it was classified as non-stressed. The threshold was chosen as 25 in this study, which corresponds to the rounded mean value of the SWICT. After the conversion, 16 samples belong to the stressed class and 22 samples belong to the non-stressed class.

4.1.3 Feature Extraction. A number of studies have identified heart rate variability (HRV) as a stress indicator and define HRV variables from both time-domain and frequency-domain [33]. For example, the standard deviation of the normal-to-normal intervals (SDNN) reflects physiological resilience against stress and it increases when HRV is large and irregular; the low-frequency (LF) band and the high-frequency (HF) band reflect the activity of sympathetic nervous system and parasympathetic nervous system respectively, and an increase in the LF/HF ratio is significantly associated with psychological stress. However, since the Cosinuss Two in-ear sensor, as many easily available heart rate sensors, does not collect the blood volume pulse signal, the aforementioned statistics cannot be computed precisely, as calculation of HRV from bpm would only result in an average rather than the detailed data expected from HRV. Therefore, in this analysis, we extracted features solely from heart rate (beats per minute). 10 summary statistics were calculated and a brief description of these statistics is listed in Table 2. After extracting these features, we computed the variance inflation factor to check for multi-collinearity and found that Median, Max and Min were highly correlated with other features. These three features were removed as they could be derived from the remaining features, leading to 7 features for further analysis. Finally, all features were standardised to have zero mean and unit variance.

Feature name	Definition	Rank
Mean	Mean value of HR	7
Std	Standard deviation of HR	1
Range_99	99% percentile of HR − 1% percentile of HR; note that outliers are removed before calculation following the 3-IQR rule [35]	2
Mean_absdiff	Mean value of lag-1 absolute difference (i.e., the absolute value of the difference between consecutive HRs)	6
Std_absdiff	Standard deviation of lag-1 absolute difference	3
Max_absdiff	Maximum of lag-1 absolute difference	5
Q75_absdiff	75% percentile of lag-1 absolute difference	4

Table 2: Description of HR features used in the first analysis and their ranks in predicting stress.

4.2 Methods

4.2.1 Feature Selection. The small sample size in this study imposed a constraint on the model complexity, and therefore, it was crucial to reduce the feature dimension by applying either dimension reduction methods or feature selection methods. We chose the latter approach since it provided better interpretability.

More specifically, feature selection means selecting a subset of features which are most informative in predicting the class. To measure feature importance, we adopted univariate ANOVA F-test statistics [23]. The F-statistic calculates the ratio of the between-group variance to the within-group variance; a higher value indicates the feature was more important in separating the class.

In this analysis, we simulated training data 1,000 times, ranked the feature importance in each round according to the F-statistic, summed the rank values over these rounds, and finally re-ordered the features in the descending order of importance. The obtained feature ranks are listed in Table 2.

4.2.2 Machine Learning Models. Five machine learning (ML) models were considered in this analysis, with different characteristics in terms of accuracy, interpretability and computational complexity.

Logistic Regression (LR) and Ridge LR: LR is a classical classification model which returns both the predicted class and the predicted probability. Considering the feature-to-sample ratio was relatively high in this data set, we also tested Ridge LR, which adds L_2-penalisation to the LR classifier.

Decision Trees (DT) and Random Forest (RF): DT is one of the most interpretable methods as it generates a set of simple rules for classifying a sample. However, the method generally has high variance and thus RF (effectively an ensemble of DTs) is often used.

Support Vector Machines (SVM): SVM is particularly useful for classifying non-linearly separated data owing to the use of kernel functions, using the radial basis function (RBF) kernel here.

Apart from the LR classifier, all methods include some hyperparameters. For Ridge LR, we fixed the cost of the L_2 penalty to be 10. Hyperparameters in other methods were set to be default as in Python `sklearn` library. Another hyperparameter is the number of features used to train the model. This value was determined by employing 10-fold cross-validation on the training data.

When applying the above methods to our data set, one issue needed to be addressed – class imbalance. Our data set contained more samples from the non-stressed class than from the stressed class, meaning that the classifier trained on it may be biased. In the worst case, it would always predict the majority class. To this end, we followed the idea of cost-sensitive learning to assign unequal costs on misclassified samples [19]. In other words, the classifier will be penalised more when classifying a stressed sample as non-stressed than classifying a non-stressed sample as stressed.

4.3 Results

4.3.1 Performance Metrics. Due to the existence of imbalanced classes, instead of accuracy, we chose to use AUC (Area Under the receiver operating characteristic Curve) as the evaluation metric. AUC ranges between zero and one and a larger value indicates better classification performance. In addition, we report sensitivity and specificity to understand how well the method predicts each class. The higher the sensitivity (specificity, resp.), the better the stress (non-stress, resp.) class is classified.

We randomly sampled 80% of the data to form a training set and used the remaining 20% as the test set. This was repeated 100 times and the average test AUC, sensitivity and specificity were reported.

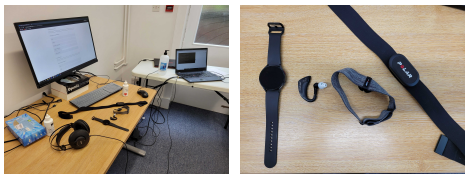
4.3.2 Performance of Machine Learning Methods. Table 3 lists the AUC score, sensitivity and specificity of different methods on test data, as well as the number of features used to build the model. LR classifier, though simple, achieved the highest AUC of 0.758 on this data set. Its sensitivity is slightly lower than the specificity, indicating the tendency towards predicting the majority class. In this optimal model, three features were selected, namely Std (the standard deviation of HR), Range_99 (the range of HR after removing outliers), and Std_absdiff (the standard deviation of lag-1 absolute difference). In addition, Table 3 reports the performance of a random classifier, which corresponds to the LR classifier trained on the same set of features but with random labels. The large gap in the AUC, sensitivity and specificity between the LR classifier and the random classifier demonstrated the predictive capability of the selected features.

4.3.3 Discussion. This analysis investigated the use of ML methods on HR data for detecting stress in real-time. Results suggest that we could achieve a relatively good detection performance solely based on HR of a short duration of one minute. Note that the data collected from the unobtrusive, comfortable sensor of Cosinuss is only HR (beats per minute), rather than more informative signals such as blood volume pulse, making it much harder to build an accurate classification model. Moreover, the data quality was a concern as participants may not wear the sensor appropriately during the entire test. We removed some samples with unreasonable heart rates based on visual inspection, however, the remaining samples analysed in this study may still be imprecise and contain noise.

Another major challenge in this analysis was the small sample size. The selected model was relatively simple, using very few features. However, adding more features and/or employing more advanced models, such as introducing non-linear kernels and ensembles, can cause severe overfitting. The small sample size also posed a difficulty for choosing the optimal set of hyperparameters.

Method	#Feat.	AUC	Sensitivity	Specificity
LR	3	0.758 (0.144)	0.698 (0.258)	0.818 (0.170)
Ridge	2	0.749 (0.153)	0.656 (0.269)	0.842 (0.175)
DT	6	0.629 (0.181)	0.546 (0.311)	0.713 (0.216)
RF	6	0.645 (0.156)	0.547 (0.277)	0.742 (0.199)
SVM	3	0.647 (0.152)	0.530 (0.285)	0.765 (0.185)
Random	3	0.497 (0.226)	0.472 (0.326)	0.523 (0.305)

Table 3: Classification performance of ML methods. Mean AUC, sensitivity and specificity averaged over 100 iterations are reported with their standard deviation in brackets.



(a) Experimental Setup. **(b) Heart rate monitors.**

Figure 2: Experimental Setup and Devices of Experiment 2. Heart Rate Monitors: Samsung Galaxy Watch4, Cosinuss One, Polar OH1, Polar H10 (from left to right)

5 EXPERIMENT 2

The second experiment was based on experiment 1, but with some minor changes in order to investigate performance of the machine learning model with different participants and devices in real-time. The changes include number of stress test components and the used devices and are detailed in each relevant subsection.

5.1 Experimental Setup

This experiment (see experimental setup in Figure 2(a)) only utilised two variable stressors of the MMST to induce stress, the PASAT and the SCWIT tests, and one constant stressor, the increasing white noise. These stressors were the most effective in experiment 1 and there were concerns that the negative images only seemed to trigger a response if participants had a traumatic event in relation to the picture. The monetary incentive was also reported to have no effect on the recruited age group.

5.2 Apparatus

The tests were conducted on a similar set-up to the previous experiment: a standard laptop with an Asus screen, a mouse for the participants to use for the PASAT. As well as the Cosinuss One[5] in-ear sensor (the predecessor of the Cosinuss Two), the participants wore a Samsung Galaxy Watch4 smartwatch[6], a Polar OH1+[7] arm strap heart rate monitor and a Polar H10[8] chest strap heart rate monitor (see Figure 2(b)). Conductive gel was used for the Polar H10, which was introduced to the study despite its more elaborate setup, compared to the other devices, to establish the ground truth for the heart rate data, as the performance of its single-lead ECG states was found to be comparable to multiple-lead ECG data [16, 30]. Smartwatch, Cosinuss and Polar OH1 used Optical Heart Rate Sensors. A pair of over-ear headphones were used to present the white noise stressor, ensuring comfortable fitting over potential hearing aids and the Cosinuss device. All devices but the smartwatch used Bluetooth to directly send the heart rate to the computer where it was logged in real-time. The smartwatch used a self-developed app that sent heart rate to the computer via WiFi (connected via mobile hotspot), which was also logged on the computer.

5.3 Participants

16 older adults (60+) were newly recruited for this experiment (10 female, 6 male). The participants were between 63 and 84 years (mean: 70.19, STDV: 7.12) and all but one right handed. The group were healthy with the same limitations as experiment 1, except allowing participants with hearing impairments with hearing aids, as long as they were only worn in a way that allowed for the Cosinuss device to be worn efficiently as well (2). Participants all used the right hand to interact with the computer mouse.

5.4 Experimental Protocol

The participants were provided with the information sheet for this experiment and a consent form. Upon completion, they filled out the STAIQ as in experiment 1. Participants who were suitable for the experiment then trialled the stress tests in the same order they would experience them to ensure they knew the test process. Participants were then equipped with heart rate sensors. The watch was placed on the wrist of the hand not using the mouse (left) to limit motion in case it affected the heart rate reading. The Cosinuss was placed in the opposite ear to the watch unless the participant required it in the other ear (for instance hearing aid being present). This switch was necessary for two participants (P02 and P08), one of which wore the Cosinuss in the same ear as the hearing aid. The participants were allowed to fit the chest strap themselves with fitting assistance from the experimenter when required. When the sensors were placed, the participants underwent a 5min biophysical data recording phase to collect a baseline measurement. Afterwards, the first of the two stress tests began, in counterbalanced order. The participants filled in a 100-point Likert scale rating their stress before the stress test began. After the test another stress rating was collected and participants then started a 5min calming phase. The participants could then take a further break or move on to the next test with the same set-up. During the stress tests, the machine learning algorithm was extracting and analysing the previous 1min of heart rate data, presenting a real-time stress detection every 10s to the experimenter by presenting a 1 when stress was detected and 0 otherwise. The 10s calculation was chosen to present usable continuous feedback without overloading the system, but could be adjusted to fit other time frames. Our college's Ethics Committee approved the study design.

PartID	P01	P02	P03	P04	P05	P06	P07	P08
PASAT	10	24	-21	8	**51**	0	-9	**65**
SWICT	8	**45**	-1	**43**	25	0	0	**38**
PartID	P09	P10	P11	P12	P13	P14	P15	P16
PASAT	**41**	**43**	0	-5	**28**	6	10	**35**
SWICT	**43**	**29**	14	1	13	21	8	20

Table 4: Stress Rating of Exp. 2, showing the differences between the Likert rating (0-100) before and after each test; rating differences of more than 25 are being shown in bold.

5.5 Results

We will again describe the collected data, such as stress rating and descriptive heart rate data, in this section and include the evaluation of the accuracy of the original stress model on the new data from all devices. As the errors participants made in the stress test were not relevant for us, we did not evaluate these.

5.5.1 Stress Ratings. The stress ratings are presented in Table 4 showing the differences between the subjective stress rating before and after the stress test on 100-point Likert scales. 12 reported stress events (bold values in Table 4) occurred in this experiment, defined as having a difference of at least 25 between pre- and post-ratings. There were again slightly more stress events for the PASAT (mean: 17.88, STDV: 24.04) than the SWICT (mean: 19.19, STDV: 16.49). 8 participants did not show any stress events during the experiment.

5.5.2 Descriptive Heart Rate Data. The heart rate data for each participant during the different stress tests can be seen for all devices in Table 5. While the Polar H10 worked well throughout the experiment, there were difficulties with the other devices. The Polar OH1 worked well for all but one participant (P14), where the data differed considerably from the other devices, showing on average almost half the heart rate (labelled **(V)** in Table 5). The Samsung Galaxy Watch4 stopped recording for some participants, leading to missing watch data for eight stress tests over five participants. The Cosinuss data had several instances of 0 loggings over several minutes (marked with **(E)** in Table 5), combined with considerably different heart rate means in some cases (labelled **(V)** in Table 5). The heart rate data collected from the different sensors was fairly consistent in some instances, see P02 in Figure 3(a), but also showed some inconsistent data over devices, as P03 in Figure 3(b). The most variability was seen for the Cosinuss sensor (blue line), which, apart from showing some very different heart rates for some time frames,

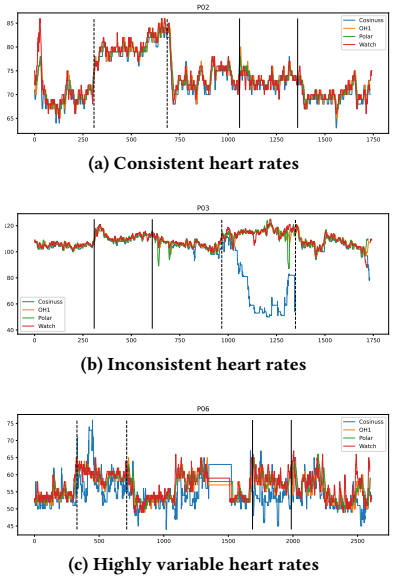
[5]https://www.cosinuss.com/en/products/in-ear-sensors/one/ (accessed 27/01/2023)
[6]https://www.samsung.com/global/galaxy/galaxy-watch4/specs/ (accessed 27/01/2023)
[7]https://www.polar.com/en/sensors/oh1-optical-heart-rate-sensor/ (accessed 27/01/2023)
[8]https://www.polar.com/en/sensors/h10-heart-rate-sensor (accessed 27/01/2023)

ID	PASAT			
	Cosinuss	Polar OH1	Polar H10	Watch
P01	91.65 (9.11)	93.49 (8.40)	94.25 (7.99)	93.75 (7.35)
P02	73.02 (1.62)	73.57 (1.49)	73.52 (1.30)	73.10 (1.40)
P03	110.90 (2.92)	111.67 (2.80)	111.88 (3.11)	112.11 (2.99)
P04	72.05 (1.15)	72.51 (0.83)	72.51 (0.81)	N/A
P05	**79.72 (1.25)**	**80.62(1.34)**	**80.71 (1.23)**	
P06	53.33 (3.42)	58.12 (2.17)	58.59 (1.94)	58.80 (2.34)
P07	67.51 (4.56)	71.07 (1.68)	71.02 (1.43)	71.56 (1.99)
P08	**74.38 (2.38)**	**74.89 (2.24)**	**75.16 (2.18)**	N/A
P09	**89.28 (2.07)**	**90.22 (2.40)**	**90.20 (2.44)**	**90.69 (2.51)**
P10	**81.96 (3.19)**	**82.45 (3.39)**	**83.62 (3.30)**	**82.33 (3.49)**
P11	77.66 (6.17)	78.16 (6.11)	78.25 (6.12)	78.65 (6.27)
P12	80.70 (1.97)	82.22 (1.61)	82.20 (1.41)	82.61 (1.69)
P13	**82.91 (2.04)**	**84.06 (2.16)**	**84.20 (1.98)**	**85.19 (2.74)**
P14	76.22 (4.92) **(E)**	42.23 (2.46) **(V)**	80.57 (4.90)	N/A
P15	70.24 (3.28)	70.90 (3.29)	70.98 (3.10)	N/A
P16	**77.12 (1.13)**	**78.02 (1.23)**	**77.78 (1.27)**	**78.99 (1.76)**
ID	SWICT			
	Cosinuss	Polar OH1	Polar H10	Watch
P01	58.44 (4.47) **(E)**	84.57 (4.22)	84.61 (4.05)	84.73 (4.07)
P02	**79.38 (1.91)**	**80.10 (2.12)**	**80.08 (2.06)**	**78.98 (1.78)**
P03	73.85 (21.43) **(V)**	114.34 (2.97)	114.30 (4.93)	115.05 (3.58)
P04	**73.86 (1.00)**	**74.25 (1.02)**	**74.26 (0.87)**	**74.58 (0.98)**
P05	**78.45 (1.80)**	**79.28 (1.88)**	**79.34 (1.76)**	**78.88 (2.24)**
P06	58.50 (5.46)	59.91 (1.78)	59.91 (1.51)	60.31 (1.72)
P07	56.35 (4.15) **(E)**	68.46 (1.19)	68.41 (1.05)	N/A
P08	**78.13 (1.23)**	**79.02 (1.04)**	**79.20 (0.83)**	N/A
P09	**83.47 (1.80)**	**84.47 (1.62)**	**84.46 (1.61)**	**84.91 (1.65)**
P10	**74.97 (9.86)**	**82.78 (2.00)**	**83.18 (2.60)**	**82.79 (2.05)**
P11	71.73 (2.88)	72.52 (2.73)	72.56 (2.58)	73.02 (0.82)
P12	87.88 (2.86)	88.98 (2.79)	89.01 (2.80)	89.39 (2.81)
P13	78.54 (2.14)	79.40 (2.20)	79.63 (2.62)	82.94 (3.83)
P14	N/A	42.25 (0.84) **(V)**	84.57 (2.55)	N/A
P15	69.53 (4.57)	70.72 (4.68)	70.73 (4.40)	N/A
P16	80.93 (3.92)	81.93 (4.01)	81.89 (4.02)	82.62 (4.25)

Table 5: Heart Rate Data of Exp. 2, showing the mean heart rate and standard deviation (in brackets) for each stress test and device; (E) is added if there were errors in the heart rate collection, i.e. 0 was logged during the test; (V) is added if a heart rate varied considerably from the other sensors; values with a stress rating of over 25 are being shown in bold font.

as seen in Figure 3(b), could also show considerably higher variation during the whole data collection, see P06 in Figure 3(c).

5.5.3 Observations. Six participants had to wear the Samsung Galaxy Watch4 with the watch face on the wrist for the heart rate to be picked up (P01, P02, P04, P08, P11, P13) and the heart rate was not logged for some of these participants (see Table 5). This could be due to the watch accidentally closing the app when the arm was moved, but that was, unfortunately, not checked and watch data for some other participants was also missing, so there might have been another issue, which we cannot recreate.
Some participants showed and reported problems with distinguishing the colours yellow, orange and red during the SWICT test, even though they were introduced to all the colours before. As we have



(a) Consistent heart rates

(b) Inconsistent heart rates

(c) Highly variable heart rates

Figure 3: Heart rate signals from four sensors. PASAT and SWICT experiments are indicated by solid and dashed lines, respectively. Blue: Cosinus, orange Polar OH1, Green: Polar H10, Red: Samsung Galaxy Watch

not evaluated the test results, we do not expect this to have an influence on our results, but it could potentially have increased the stress of these participants during the experiment. One participant mentioned that the presence of the experimenter was experienced as a calming influence and another participant mentioned that the white noise was calming rather than stressful.

5.5.4 Performance of Pre-trained Model for Stress Detection. The purpose of this analysis was to evaluate the effectiveness of the model learned from the first experiment in detecting stress for new participants and with different devices.
Heart rate data and stress rating from experiment 2 were processed in the same way as in the experiment 1, with details explained in Sections 4.1.1 and 4.1.2. After removing the erroneous measurement from participant 14, there were 30 samples available for analysis on Cosinuss, Polar OH1 and H10. Among them, 12 samples belonged to the stressed and 18 samples belonged to the non-stressed class. For Watch data, there were more missing data as shown in Table 5 and only 24 samples could be used for analysis, with 10 and 14 samples from the stressed and non-stressed class, respectively. Feature extraction followed the methods described in 4.1.3.

The classification results of applying the pre-trained LR classifier, i.e. the classifier trained on data from the experiment 1, to data from the second experiment showed reduced performance. Recall that in the first analysis, the model achieved an AUC of 0.758. However, in this analysis, the performance on Cosinuss data was only 0.514 (sensitivity: 0.250, specificity: 0.778). The low sensitivity suggested that the majority of stressed samples were misclassified as non-stressed. When applying the model to Polar OH1 (AUC: 0.597, sensitivity: 0.250, specificity: 0.944) and Polar H10 (AUC: 0.569,

sensitivity: 0.250, specificity: 0.889 data, the results were slightly better, but still close to 0.5 (a random classifier). The performance on Watch data is even worse (AUC: 0.407, sensitivity: 0.100, specificity: 0.714). Note that the result on Watch is not directly comparable with those from Cosinuss, Polar OH1 and Polar H10 as there are less samples in the Watch data set.

5.5.5 Discussion. The adapted Mannheim Multicomponent Stress Test with only one constant stressor and two variable stressor led to 12 stress events in 16 participants, only slightly less than in the first experiment. Both tests did not seem to induce stress in some participants. Previous research suggested that singing without music was a very effective task to induce stress and could be adapted for further experiments, if a more effective stress induction is needed. Additionally, the white noise was described as calming rather than stressful, so it might be more stressful to use different kind of noise such as heavy construction noise.

The heart rate data showed considerable differences between the devices. Some devices did not reliably collect the data during the experiment, especially the watch and the Cosinuss were unstable. The Cosinuss showed overall more variability in the heart rate data than the other three devices. As the previous study also used a Cosinuss device (albeit a different version), this could have heavily influenced the stress detection.

The analysis investigating the applicability of classification models across participants and sensors and evaluating the effectiveness of different sensors for stress detection suggested that the model learned from one cohort of participants did not generalise to another cohort of participants. In fact, this finding should not come as a surprise given the large discrepancy in the variability of heart rates between the two experiments as shown in Tables 1 and 5 and that variability was the key predictor of stress.

Results also suggested that devices have a large impact on stress detection. Again, due to the discrepancy in data characteristics between sensors, both pre-trained and re-trained models on Cosinuss data failed to generalise to the other three data sets. A possible reason for this could be that models trained on Cosinuss and Watch took advantage of noise or artificial variability in the data sets.

6 OVERALL DISCUSSION

In this section we will discuss both experiments and the stress detection models. We will start of with discussing the limitations of the research and then move to the MMST, followed by the stress detection system and heart rate devices.

6.1 Limitations

This research investigated the usefulness of heart beats per minute for stress detection of older adults. We collected heart rate data from older adults in several stress inducing tests. These tests have been chosen according to previous research, but different tests worked better on different individuals and they did not induce stress in all adults. Because of the specific participant health requirements, many interested older adults could not take part in the studies, decreasing further the already hard to recruit target group of older adults and the number of participants can have an influence on the machine learning model accuracy. Still, participant numbers in our experiments were higher than in some of the prior research [5, 14]. In addition, the heart rate devices did not always collect data reliably, decreasing the available data for model training. The data was collected while participants were sitting, so the results would have to be adjusted to fit situations in which users would be moving around, as this would have an influence on the heart rate that has not been considered in this research. Applications like tele-working or within an office setting would be potential suitable use cases for the research presented in this paper. We collected self-reported stress ratings to determine the stress level of the participants and used these as basis for our stress detection model. These ratings could be biased and not correspond with the bodily reaction, as humans might interpret stress differently and, therefore, rate it differently. We chose to use the 1min heart rate data from just before the end of the stress tests to built our model, arguing that this would be the time in which participants would the most stressed, without having any external influence changing the heart rate. We cannot be sure that this is the time in which participants experience the most stress and the data might not be optimal.

6.2 Mannheim Multicomponent Stress Test

The Multicomponent Stress Test (MMST) has been used in this research to induce stress safely in older adults. The original MMST used three variable stressors and two constant stressors to induce stress. The two constant stressors used in the first experiment were noise and loss of money for mistakes. We found that the older adults in the first experiment did not care about the monetary incentive at all, some even refusing to receive any money for their participation. We, therefore, decided not to use this test in the second experiment. Further, we found the affective images (AffIm) to be less effective with this specific population group. Only one stress event was triggered by AffIm, and the participant later commented on this having been due to childhood trauma related to the triggering picture, which could have a longer-lasting effect than intended by these kinds of tests. In the second experiment we, therefore, only used the more effective stress inducing tests: the Stroop Colour Word Interference Test (SCWIT) and the Paced Auditory Serial Addition Test (PASAT). One of these tests induced stress in most adults in the first experiment. Only two adults did not report stress, defined as a difference of at least 25 between 100-point Likert scale ratings before and after the test. In the second experiment, half of the participants did not rate any test as stressful within our definition. The two experiments were conducted in different countries and cities and participants were recruited differently: the first experiment was conducted in a small city with no university close by and participants were recruited from that smaller, local community, while the second experiment was conducted in a large city with several universities and participants were also contacted through email lists of already more research experienced groups of older adults. This could have an impact on how they reacted to the tests, as the testing situation itself, as well as the use of technology, could potentially increase stress in people with no prior experience. Researchers depending on inducing stress in older adults should skip the use of affective images and focus on finding alternative sources of potential stress inducting activities. Singing has been named as an effective stress induction test [14]; this could be tested on older adults in future work to see if stress could be induced more equally.

6.3 Stress Detection and Heart Rate Devices

After the first experiment using the Cosinuss Two heart rate sensor, we built a machine learning model to use logged heart beats per minute to detect stress in real-time, based on the self-reported stress ratings collected before and after each stress test. Our model showed a promising classification performance of 76%, evaluating beats per minute over one minute of heart rate data. We then applied this detection model in real-time, with stress values being presented every 10s, during a second experiment, in which we collected heart rate data with four different devices: the Cosinuss One (predecessor of the Cosinuss Two used in the first experiment), the Samsung Galaxy Watch4 (running a self-developed app), and the chest-worn Polar H10 and arm-worn Polar OH1.

The performance of the model dropped to 51% for the same device and between 60% and 41% for the other devices using the same model. This may be a consequence of less variability in the heart rate data in the second experiment than in the first experiment, and additionally all other three devices collected less variable heart rate data than Cosinuss did. As the Cosinuss data is the only one showing high variation in the data, it is possible that an unstable element in the device's heart rate sensor rather than the actual heart rate led to those variations and resulted in better classification performance.

In summary, current analyses showed that a model built on data from one device may be ineffective for data collected from a different device. Heart beats per minute alone seem to be insufficient to provide enough information for reliable stress detection in older adults. However, it is unclear whether a model trained on more data can generalise well to different cohorts of participants and/or devices, and thus more research with more participants is needed to verify this.Previous research has often discussed successful machine learning models based on a low number of participants, but as our testing a similar promising model in a second experiment showed: the high accuracy does not necessarily translate to newly acquired data and without a similar evaluation study these discrepancies would not have been found.

7 CONCLUSIONS

We collected heart rate data from older adults to build a real-time stress detection model on heart beats per minute, as this data can be easily obtained with a plethora of single devices. Older adults are apt to turn from technology use if the technology becomes inconvenient, so we wanted to explore the effectiveness of using one of the most easily available physiological measures. We evaluated how well the Mannheim Multicomponent Stress Test can be used with older adults and found that some of the classical stressors used in the test did not induce stress with this cohort. A reduced stress test did still induce stress, but not in all participants. Different types of stressors need to be investigated to ensure stress induction in older adults. The stress detection model built on the data of the first experiment showed promising classification performance of 76% (AUC), but when used in the second experiment this dropped to 51% (AUC). This evaluation step of machine learning models in a second experiment has often been omitted in previous research, but our findings show that high calculated accuracy on one set of data does not necessarily translate well. Other devices have been tested in the same experiment, and their prediction performance was comparably low. All but one device had some inaccuracy or data loss during the heart rate collection and the three additional devices used in the second study differed in variability from the device used in both experiments, leading to low prediction performance. This would lead to the conclusion that heart beats per minute alone might not be enough to build a successful stress detection model in this form and that models for specific devices might not easily be used with other heart rate sensors. These findings show the need to find alternative, easily obtainable physiological data which could be used in addition to heart rate used to detect stress in older adults.

ACKNOWLEDGMENTS

This research was part of the RadioMe project (EPSRC (EP/S026991/1)).

REFERENCES

[1] Tim Adams, Sophie Wagner, Melanie Baldinger, Incinur Zellhuber, Michael Weber, Daniel Nass, and Rainer Surges. 2022. Accurate detection of heart rate using in-ear photoplethysmography in a clinical setting. *Frontiers in Digital Health* 4, August (2022), 1–10. https://doi.org/10.3389/fdgth.2022.909519

[2] Md. Abu Baker Siddique Akhonda, Shaon Foorkanul Islam, Ahmed Shehab Khan, Fariha Ahmed, and Mostafizur Rahman. 2014. Stress Detection of Computer User in Office like Working Environment Using Neural Network. *17th International Conference on Computer and Information Technology (ICCIT) Stress* (2014), 174–179.

[3] Anjana Bali and Amteshwar Singh Jaggi. 2015. Clinical experimental stress studies: Methods and assessment. *Reviews in the Neurosciences* 26, 5 (2015), 555–579. https://doi.org/10.1515/revneuro-2015-0004

[4] Hege Bøen, Odd Steffen Dalgard, and Espen Bjertness. 2012. The importance of social support in the associations between psychological distress and somatic health problems and socio-economic factors among older adults living at home: a cross sectional study. *BMC geriatrics* 12 (2012), 27. https://doi.org/10.1186/1471-2318-12-27

[5] Sansanee Boonnithi and Sukanya Phongsuphap. 2011. Comparison of heart rate variability measures for mental stress detection. *Computing in Cardiology* 38 (2011), 85–88.

[6] Yekta Said Can, Bert Arnrich, and Cem Ersoy. 2019. Stress detection in daily life scenarios using smart phones and wearable sensors: A survey. *Journal of Biomedical Informatics* 92, February (2019), 103139. https://doi.org/10.1016/j.jbi.2019.103139

[7] R. Castaldo, L. Montesinos, P. Melillo, S. Massaro, and L.Pecchia. 2018. To What Extent Can We Shorten HRV Analysis in Wearable Sensing? A Case Study on Mental Stress Detection. *IFMBE Proceedings* 65 (2018), 1089–1090. https://doi.org/10.1007/978-981-10-5122-7{_}161

[8] Waite L. Cornwell, E.Y. 2009. Social Disconnectedness, Perceived Isolation, and Health among Older Adults. *Journal Health Social Behaviour* 50, 1 (2009), 31–48. https://doi.org/nihms-133647

[9] Raritan Costin, Cristian Rotariu, and Alexandru Pasarica. 2012. Mental stress detection using heart rate variability and morphologic variability of EeG signals. *EPE 2012 - Proceedings of the 2012 International Conference and Exposition on Electrical and Power Engineering* Epe (2012), 591–596. https://doi.org/10.1109/ICEPE.2012.6463870

[10] Elise S. Dan-Glauser and Klaus R. Scherer. 2011. The Geneva affective picture database (GAPED): A new 730-picture database focusing on valence and normative significance. *Behavior Research Methods* 43, 2 (2011), 468–477. https://doi.org/10.3758/s13428-011-0064-1

[11] Franca Delmastro, Flavio Di Martino, and Cristina Dolciotti. 2020. Cognitive Training and Stress Detection in MCI Frail Older People through Wearable Sensors and Machine Learning. *IEEE Access* 8, Mci (2020), 65573–65590. https://doi.org/10.1109/ACCESS.2020.2985301

[12] Michael C. Diehr, Robert K. Heaton, Walden Miller, and Igor Grant. 1998. The Paced Auditory Serial Additiona Task (PASAT): Norms for Age, Education, and Ethnicity. *Assessment* (1998).

[13] Kathryn Dreyer, Adam Steventon, Rebecca Fisher, and Sarah R. Deeny. 2018. The association between living alone and health care utilisation in older adults: A retrospective cohort study of electronic health records from a London general practice. *BMC Geriatrics* 18, 1 (2018), 15–17. https://doi.org/10.1186/s12877-018-0939-4

[14] Begum Egilmez, Emirhan Poyraz, Wenting Zhou, Gokhan Memik, Peter Dinda, and Nabil Alshurafa. 2017. UStress: Understanding college student subjective stress using wrist-based passive sensing. *2017 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2017* (2017), 673–678. https://doi.org/10.1109/PERCOMW.2017.7917644

[15] Barbara L Fredrickson and Daniel Kahneman. 1993. Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology* 65, 1 (1993), 45.

[16] Rahel Gilgen-Ammann, Theresa Schweizer, and Thomas Wyss. 2019. RR interval signal quality of a heart rate monitor and an ECG Holter at rest and during exercise. *European Journal of Applied Physiology* 119, 7 (2019), 1525–1532. https://doi.org/10.1007/s00421-019-04142-5

[17] Adam M. Goodman, Jeffrey S. Katz, and Michael N. Dretsch. 2016. Military Affective Picture System (MAPS): A new emotion-based stimuli set for assessing emotional processing in military populations. *Journal of Behavior Therapy and Experimental Psychiatry* 50 (2016), 152–161. https://doi.org/10.1016/j.jbtep.2015.07.006

[18] Anke Haberkamp, Julia Anna Glombiewski, Filipp Schmidt, and Antonia Barke. 2017. The DIsgust-RelaTed-Images (DIRTI) database: Validation of a novel standardized set of disgust pictures. *Behaviour Research and Therapy* 89 (2017), 86–94. https://doi.org/10.1016/j.brat.2016.11.010

[19] Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* 21, 9 (2009), 1263–1284.

[20] Karen Hovsepian, Mustafa Al'absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. 2015. CStress: Towards a gold standard for continuous stress assessment in the mobile environment. *UbiComp 2015 - Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2015), 493–504. https://doi.org/10.1145/2750858.2807526

[21] Bettina S. Husebo, Hannah L. Heintz, Line I. Berge, Praise Owoyemi, Aniqa T. Rahman, and Ipsit V. Vahia. 2020. Sensing technology to facilitate behavioral and psychological symptoms and to monitor treatment response in people with dementia: A systematic review. *Frontiers in Pharmacology* 10, February (2020), 1–13. https://doi.org/10.3389/fphar.2019.01699

[22] Tatyana Kolotylova, Mandy Koschke, Karl Jürgen Bär, Ulrich Ebner-Priemer, Nikolaus Kleindienst, Martin Bohus, and Christian Schmahl. 2010. Entwicklung des mannheimer multikomponenten-stress-test (MMST). *PPmP Psychotherapie Psychosomatik Medizinische Psychologie* 60, 2 (2010), 64–72. https://doi.org/10.1055/s-0028-1103297

[23] Max Kuhn and Kjell Johnson. 2019. *Feature engineering and selection: A practical approach for predictive models*. CRC Press.

[24] Kari Kvaal, Ingun Ulstein, Inger Hilde Nordhus, and Knut Engedal. 2005. The Spielberger State-Trait Anxiety Inventory (STAI): the state scale in detecting mental disorders in geriatric patients. *INTERNATIONAL JOURNAL OF GERIATRIC PSYCHIATRY* (2005). https://doi.org/10.1002/gps.1330

[25] S. J. Lupien, F. Maheu, M. Tu, A. Fiocco, and T. E. Schramek. 2007. The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition. *Brain and Cognition* 65, 3 (2007), 209–237. https://doi.org/10.1016/j.bandc.2007.02.007

[26] Jesus Minguillon, Eduardo Perez, Miguel Angel Lopez-Gordo, Francisco Pelayo, and Maria Jose Sanchez-Carrion. 2018. Portable system for real-time detection of stress level. *Sensors (Switzerland)* 18, 8 (2018), 1–15. https://doi.org/10.3390/s18082504

[27] Tracy L. Mitzner, Julie B. Boron, Cara Bailey Fausset, Anne E. Adams, Neil Charness, Sara J. Czaja, Katinka Dijkstra, Arthur D. Fisk, Wendy A. Rogers, and Joseph Sharit. 2010. Older adults talk technology: Technology usage and attitudes. *Computers in Human Behavior* 26, 6 (2010), 1710–1721. https://doi.org/10.1016/j.chb.2010.06.020

[28] Martin J. Prince, Fan Wu, Yanfei Guo, Luis M. Gutierrez Robledo, Martin O'Donnell, Richard Sullivan, and Salim Yusuf. 2015. The burden of disease in older people and implications for health policy and practice. *The Lancet* 385, 9967 (2015), 549–562. https://doi.org/10.1016/S0140-6736(14)61347-7

[29] Hillol Sarker, Matthew Tyburskir, M. D.Mahbubur Rahman, Karen Hovsepian, Moushumi Sharmin, David H. Epsteinr, Kenzie L. Prestonr, C. Debra Furr-Holden, Adam Milam, Inbal Nahum-Shani, Mustafa Al'Absi, and Santosh Kumar. 2016. Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data. *Conference on Human Factors in Computing Systems - Proceedings* (2016), 4489–4501. https://doi.org/10.1145/2858036.2858218

[30] Marcelle Schaffarczyk, Bruce Rogers, Rüdiger Reer, and Thomas Gronwald. 2022. Validity of the Polar H10 Sensor for Heart Rate Variability Analysis during Resting State and Incremental Exercise in Recreational Men and Women. *Sensors* 22, 17 (2022). https://doi.org/10.3390/s22176536

[31] Charles D Spielberger, Fernando Gonzalez-Reigosa, Angel Martinez-Urrutia, Luiz F S Natalicio, and Diana S Natalicio. 1971. The State-Trait Anxiety Inventory. *Revista Interamericana de Psicologia/Interamerican Journal of Psychology* 5, 3 \& 4 (1971).

[32] J. R. Stroop. 1935. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology* 18, 6 (1935), 643–662. https://doi.org/10.1037/h0054651

[33] Julian F Thayer, Fredrik Åhs, Mats Fredrikson, John J Sollers III, and Tor D Wager. 2012. A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health. *Neuroscience & Biobehavioral Reviews* 36, 2 (2012), 747–756.

[34] S. Tivatansakul and M. Ohkura. 2015. Improvement of emotional healthcare system with stress detection from ECG signal. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* 2015-Novem (2015), 6792–6795. https://doi.org/10.1109/EMBC.2015.7319953

[35] John W Tukey et al. 1977. *Exploratory Data Analysis*. Vol. 2. Reading, MA.

[36] Kenneth J Turner, Brian O Neill, Gary Cornelius, and Evan H Magill. 2017. Predicting Emotional Dysregulation. *Technical Report CSM-200* September (2017).

[37] Jacqueline Wijsman, Bernard Grundlehner, Hao Liu, Hermie Hermens, and Julien Penders. 2011. Towards mental stress detection using wearable physiological sensors. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* (2011), 1798–1801. https://doi.org/10.1109/IEMBS.2011.6090512

[38] Janine L. Wiles, Annette Leibing, Nancy Guberman, Jeanne Reeve, and Ruth E.S. Allen. 2012. The meaning of "aging in place" to older people. *Gerontologist* 52, 3 (2012), 357–366. https://doi.org/10.1093/geront/gnr098

[39] Pamela Zontone, Antonio Affanni, Alessandro Piras, and Roberto Rinaldo. 2022. Exploring Physiological Signal Responses to Traffic-Related Stress in Simulated Driving †. *Sensors* (2022).