Taylor & Francis
Taylor & Francis Group

IISE Transactions on Healthcare Systems Engineering

ISSN: 2472-5579 (Print) 2472-5587 (Online) Journal homepage: https://www.tandfonline.com/loi/uhse21

Continuous monitoring and detection of post-traumatic stress disorder (PTSD) triggers among veterans: A supervised machine learning approach

Anthony D. McDonald, Farzan Sasangohar, Ashish Jatav & Arjun H. Rao

To cite this article: Anthony D. McDonald, Farzan Sasangohar, Ashish Jatav & Arjun H. Rao (2019) Continuous monitoring and detection of post-traumatic stress disorder (PTSD) triggers among veterans: A supervised machine learning approach, IISE Transactions on Healthcare Systems Engineering, 9:3, 201-211, DOI: 10.1080/24725579.2019.1583703

To link to this article: https://doi.org/10.1080/24725579.2019.1583703

IISE TRANSACTIONS ON HEALTHCARE SYSTEMS ENGINEERING
2019, VOL. 9, NO. 3, 201–211
https://doi.org/10.1080/24725579.2019.1583703

Taylor & Francis
Taylor & Francis Group

Continuous monitoring and detection of post-traumatic stress disorder (PTSD) triggers among veterans: A supervised machine learning approach

Anthony D. McDonald, Farzan Sasangohar, Ashish Jatav, and Arjun H. Rao

Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX, USA

ABSTRACT
Post-traumatic stress disorder (PTSD) is a prevalent mental health condition among United States combat veterans, associated with high incidence of suicide and substance abuse. While PTSD treatments exist, such methods are limited to in-person therapy sessions and medications. Tools and technologies to monitor patients continuously, especially between sessions, are largely absent. This article documents efforts to develop predictive algorithms that utilize real-time heart rate data, collected using commercial off-the-shelf wearable sensors, to detect the onset of PTSD triggers. The heart rate data, pre-processed with a Kalman filter imputation approach to resolve missing data, were used to train five algorithms: decision tree, support vector machine, random forest, neural network, and convolutional neural network. Prediction performance was assessed with the Area Under the receiver operating characteristic Curve (AUC). The convolutional neural network, support vector machine, and random forests had the highest AUC and significantly outperformed a random classifier. Further analysis of the heart rate data and predictions suggest that the algorithms associate an increase in heart rate with PTSD trigger onset. While work is needed to enhance algorithm performance and robustness, these results suggest that wearable monitoring technology for PTSD symptom mitigation is an achievable goal in the near future.

KEYWORDS
Convolutional neural networks; human physiology; post-traumatic stress disorder; random forest; wearable technology

1. Introduction

Post-traumatic stress disorder (PTSD) is a prevalent mental disorder that commonly occurs after an individual has experienced a traumatic event, such as the death of loved ones, sexual violence, or serious injury (U.S. Department of Veterans Affairs, 2017). Symptoms of PTSD involve: (1) re-experience (e.g., flashbacks and nightmares); (2) avoidance of situations reminiscent of the event; (3) negative changes to belief and cognition; and (4) heightened anxiety and alertness (also known as hyperarousal) (American Psychiatric Association, 2013). While several high-risk professions, such as firefighters, police, and emergency department clinicians, are exposed to frequent traumatic events, combat veterans stand out as especially prone to extreme forms of such events. In fact, 11–20% of the veterans who served in Operations Iraqi Freedom (OIF) and Enduring Freedom (OEF), 12% of the veterans who served in the Gulf War (Desert Storm), and 15% of the veterans who served in the Vietnam War suffered from PTSD according to various studies (U.S. Department of Veterans Affairs, 2016). PTSD has been associated with high rates of suicide among combat veterans (Baker, 1984; Hendin and Haas, 1991; Lawrence *et al.*, 1985).

Individuals suffering from PTSD are treated using pharmacotherapy (medication), psychotherapy ("talk" therapy), or both. Psychotherapy, which has been shown to be more effective in treating PTSD patients (Forbes *et al.*, 2010), primarily applies prolonged exposure (PE) therapy or cognitive processing therapy (CPT). While PE attempts to help patients overcome their fears by exposing them repeatedly to situations that trigger PTSD symptoms (Foa *et al.*, 2008), CPT helps change patients' understanding and perception of traumatic events by removing the burden of blame/fault (U.S. Department of Veterans Affairs, 2018). Despite having treatment options, most approaches are limited to in-person sessions with a therapist and have no capabilities to monitor patients between sessions. Furthermore, therapy session attendance may be affected by geographic, social, financial, and temporal barriers that inhibit timely treatment of PTSD patients.

Recent technological advances have resulted in the development of mobile health (mHealth) applications (referred to as "apps"). The emergence of mHealth apps has enabled patients to monitor their symptoms, carry out therapeutic work asynchronously, communicate (and socialize) with other patients, and contact clinicians or emergency personnel in the event of crises (Rodriguez-Paras *et al.*, 2017; Sloan *et al.*, 2011; Al Ayubi *et al.*, 2014). A recent review by Rodriguez-Paras *et al.* (2017) showed that, while technologies exist for PTSD management, few applications have been directly integrated with clinical PTSD treatments, thereby making it challenging to potentially monitor PTSD patients and provide timely care and support. While there have been recent advances in sensor technology that enable real-time

CONTACT Farzan Sasangohar sasangohar@tamu.edu 3131 TAMU, College Station, TX 77433
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uhse.

Table 1. Summary of stress detection algorithms.

Reference	Domain	Features	Classifier	Ground truth
(Picard *et al.*, 2001)	General	Statistical and domain-based features for HR, SC, facial muscle tension, and breathing rate	MAP classifier k-Nearest Neighbor	Multiclass Classes of emotions
(Healey & Picard, 2005)	Driving	Statistical, spectral, and domain-based features for HR, SC, and breathing rate	Linear Discriminant Analysis	Multiclass Rest, highway driving, and city driving
(Zhai & Barreto, 2006)	General	HRV statistical features Power-spectrum based heart rate features Galvanic skin response Statistical and non-linear features Pupil diameter Body temperature	Naive Bayes Decision Tree Support Vector Machine	Binary Mental stress from normal stress
(Boonnithi *et al.*, 2011)	General	HRV statistical features HR statistical features Power spectrum-based features Sympathetic modulation index Vagal modulation index Sympathovagal balance index	Not reported	Binary Mental stress from normal stress
(Costin *et al.*, 2012)	Driving	Mean HR Mean HRV Morphological Validity from HRV	Not reported	Multiclass Normal, low, and high mental stress based on driving environment
(Holmgard *et al.*, 2013)	PTSD treatment	Statistical features of Skin conductance	Correlative analysis	Correlations with PTSD profiles and Stress Self-reports
(Bousefsaf *et al.*, 2013)	General	Heart rate variability measured by remote camera	Correlative analysis	Correlations with skin conductance measures
(Singh *et al.*, 2013)	Driving	Statistical and syntactic features of galvanic skin response and PPG Spectral features of HRV Statistical features of HRV	Neural Networks	Multiclass Relaxed, Moderate Stress, and Stressed, based on driving environment
(Melillo *et al.*, 2013)	Student exams	Non-linear HRV features based on Poincare plots and approximate entropy	Linear Discriminant Analysis	Binary Pre-exam stress and post-holiday stress
(Sano & Picard, 2013)	Naturalistic observations of general stress	Survey responses Skin conductance statistical and non-linear features Accelerometer statistical and non-linear features	Support Vector Machine k-Nearest Neighbor	Binary Subjective ratings of low and high stress
(McDuff *et al.*, 2014)	General	HR, HRV, and breathing rate-based features	Support Vector Machine Naive Bayes	Binary Mental stress induced through arithmetic and a rest state
(Zheng *et al.*, 2016)	Competitive cycling	HR statistical and frequency features Frequency and wavelet-based EEG features	Support Vector Machine k-Nearest Neighbor	Multiclass Low, medium, and high anxiety based on a composite rating
(Ciabattoni *et al.*, 2017)	General	Statistical features of HRV Statistical features of galvanic skin response Mean and maximum body temperature	k-Nearest Neighbor	Binary Mental stress from normal stress

tracking for various conditions such as stroke and diabetes (Rand *et al.*, 2009; Zanon *et al.*, 2012), to the best of the authors' knowledge, there exists no data-driven, continuous monitoring tool to predict the onset of PTSD triggers. To address this gap, work is in progress to design a veteran-centered mHealth app to detect the onset of PTSD triggers, and engage veterans in therapeutic activities and support systems (see Williams *et al.*, 2018; Khanade and Sasangohar, 2017). A central component of this app is the PTSD trigger detection algorithm. Following a review of related literature, this article documents our computational approach for detecting PTSD triggers by using supervised machine learning techniques—similar to detection of elevated stress levels—on "ground truth" physiological data collected naturalistically from veterans during several cycling events.

1.1. Prior algorithms for stress detection

The challenge of using physiological data to differentiate between levels of human stress has received significant attention in previous work. The goals of this literature vary between exploring stress in different domains (Healey and Picard, 2005; Zheng *et al.*, 2016), identifying new measurement methods (Zhai and Barreto, 2006), or validating new measures (Costin *et al.*, 2012). In the context of machine learning, these approaches can be differentiated by their domain, input features, classification algorithms, and ground truth measures. Table 1 shows a summary of prior algorithms arranged by these facets.

1.1.1. Domain

Explorations of stress detection have primarily been conducted in laboratory settings, and focused on general stresses associated with fixed tasks (Boonnithi and Phongsuphap, 2011; Ciabattoni *et al.*, 2017; McDuff *et al.*, 2014; Picard *et al.*, 2001; Zhai and Barreto, 2006). In addition to these general studies, several studies have focused on specific domains, such as driving (Costin *et al.*, 2012; Healey and Picard, 2005; Singh *et al.*, 2013), student exams (Melillo *et al.*, 2013), and competitive cycling (Zheng *et al.*, 2016). Work on PTSD has been limited; however, Holmgard *et al.* (2013) explored stress detection among PTSD patients while participants interacted with a desktop game. PTSD monitoring through physiological reactions remains a research gap, and studies in naturalistic settings are largely absent.

1.1.2. Input measures and feature generation

Stress detection literature has explored many physiological data sources and feature types. Physiological sources explored include heart rate (HR; Boonnithi *et al.*, 2011; Costin *et al.*, 2012), heart rate variability (HRV; Bousefsaf *et al.*, 2013; Zhai and Barreto, 2006), skin conductance (SC; Ciabattoni *et al.*, 2017; Healey and Picard, 2005, 2013), breathing rate (Healey and Picard, 2005; Picard *et al.*, 2001), facial muscle tension (Picard *et al.*, 2001), body temperature (Ciabattoni *et al.*, 2017; Zhai and Barreto, 2006), body movement (Sano and Picard, 2013), and brain measures through electroencephalography (EEG; Zheng *et al.*, 2016). These measures have been collected through clinical methods (Picard *et al.*, 2001), remote camera sensing (Bousefsaf *et al.*, 2013; McDuff *et al.*, 2014), and wearable technologies (Sano and Picard, 2013; Zheng *et al.*, 2016). Analyses have identified HR and HRV as promising indicators of stress (Boonnithi *et al.*, 2011; Khanade and Sasangohar, 2017).

Given that these measures often result in noisy signals, most work has applied pre-processing methods to ensure accuracy. Feature generation is a technique commonly applied to the raw data to produce a set of features for algorithm input. Features explored previously include statistical features (e.g., mean and standard deviation), spectral features (e.g., Fourier transforms), and domain specific measures (e.g., sympathetic modulation index). In addition to these feature generation methods, several studies have employed principle components analysis (PCA) to identify combinations of effective features (Sano and Picard, 2013; Zheng *et al.*, 2016). While there is some evidence supporting the effectiveness of statistical features as identifiers of stress (Boonnithi *et al.*, 2011), other studies have shown that nonlinear and spectral features are the most effective (Melillo *et al.*, 2013). Most studies employ a combination of statistical, nonlinear, and spectral features. In these cases, features selection methods are used to reduce the feature set and identify the most effective features (e.g., Picard *et al.*, 2001).

1.1.3. Classification approaches

Many classification approaches have been used to classify mental stress from normal states; 0bhowever, these approaches have not been applied to PTSD patients in naturalistic settings. Work in other domains has used Support Vector Machines (SVM; McDuff *et al.*, 2014; Sano and Picard, 2013; Zhai and Barreto, 2006; Zheng *et al.*, 2016), k-Nearest Neighbor (k-NN; Ciabattoni *et al.*, 2017; Picard *et al.*, 2001; Sano and Picard, 2013; Zheng *et al.*, 2016), Naive Bayes (McDuff *et al.*, 2014; Zhai and Barreto, 2006), Linear Discriminant Analysis (LDA; Healey and Picard, 2005; Melillo *et al.*, 2013), neural networks (Singh *et al.*, 2013), and decision trees (Zhai and Barreto, 2006). Of the current approaches, SVM is the most widely used. Comparative analyses have shown that SVM outperforms decision trees (Zhai and Barreto, 2006) and Naive Bayes classifiers (McDuff *et al.*, 2014; Zhai and Barreto, 2006), and performs comparably to k-NN (Sano and Picard, 2013); however, given the diversity in feature generation methods, domains, and ground truths used, it is difficult to generalize these findings. Two notable exclusions from the list of current approaches are Convolutional Neural Networks (CNN) and random forests. These approaches have been widely employed in other human state detection problems, such as driver impairment (McDonald *et al.*, 2014, 2017), speech recognition (Faust *et al.*, 2018), heart attack identification (Acharya *et al.*, 2017a, 2017b), and seizure detection (Acharya *et al.*, 2017c). Random forests are an ensemble extension of decision trees that incorporate bagging—iterative random selection of feature subsets (Breiman, 2001). Trained random forest models consist of a large set of decision trees, each trained on random subsets of features. Classification in a random forest model is performed by a majority vote amongst the predictions of each tree. Theoretically, this process reduces the tendency of tree-based models to overfit the training data. CNNs are a deep-learning extension of standard multi-layer neural networks that use multiple layers as filters, or convolutional layers, to focus on specific feature groups. CNNs are particularly robust for high-dimensional feature sets and time-series classification (LeCun and Bengio, 1995). This robustness is driven by the ability of CNNs to generate their own feature spaces, which has a further advantage of eliminating the need for feature engineering. One concern with the use of random forests and CNNs is to justify the added model complexity (i.e., degrees of freedom) with additional predictive performance. Previous studies addressed this issue by fitting multiple models and statistically comparing their predictive performance (e.g., Singh *et al.*, 2013).

1.1.4. Ground truth

There are two dimensions associated with ground truth in prior work: binary versus multiclass ground truth, and the definition of stress. In binary classification studies, states are defined as either stressed or resting. Multiclass studies expand this to include medium- and high-stress states. Most work has focused on binary classification (Boonnithi *et al.*, 2011; Melillo *et al.*, 2013; Sano and Picard, 2013; Zhai and

Barreto, 2006), although several approaches investigated multiclass classification (Costin *et al.*, 2012; Picard *et al.*, 2001; Zheng *et al.*, 2016). Generally speaking, multiclass classification problems are more difficult to solve and evaluate, though they may be advantageous in practice, as different stress states may benefit from different interventions. In this study, we investigate binary classification, as our focus is differentiating between acute PTSD stress triggers and normal behavior.

The definition of stress typically depends on the domain of interest. Laboratory studies have induced stress with tasks such as the Stroop test (Bousefsaf *et al.*, 2013; Zhai and Barreto, 2006). Studies in the driving domain often relate stress to driving environments such as city or highway driving (Healey and Picard, 2005). In their investigation of PTSD, Holmgard *et al.* (2013) induced stress with a supermarket shopping simulation that included potential audio and visual PTSD triggers. Naturalistic studies have typically identified stress through subjective survey responses. This analysis uses ground truth defined by real-time participant reporting operationalized through interactions (tapping) with a smartwatch. This approach is advantageous as it is contextual, is relatively non-invasive, and has high temporal accuracy. In addition, finger tapping has been shown to reduce anxiety (Bauman and Melnyk, 1994), which may improve participation.

1.1.5. Summary

Few existing stress detection algorithms focus on the PTSD domain and consider deep-learning or ensemble learning methods. In this article, we address these gaps by developing PTSD trigger detection algorithms using deep learning and ensemble techniques. The remainder of this article describes our methods used to collect the ground truth, discusses the dataset and data processing steps, presents the results of fitted algorithms, and concludes with a discussion of findings and recommendations for future work.

2. Methods

2.1. Experimental data

Data for this research were collected during bicycle riding events organized by Project Hero—a non-profit organization that is dedicated to helping veterans and first responders with PTSD. Physiological data were collected using a wearable health monitoring tool developed at the Applied Cognitive Ergonomics Lab (ACE-Lab) at Texas A&M University. The tool is an mHealth app designed for Android Wear devices, which will be programmed with the detection algorithm described in this article to learn and respond to the wearer's PTSD-specific physiological cues. The device helps detect changes in heart rate not associated with athletics or normal activities and interacts with the wearer to help manage the onset of a PTSD trigger or contact the wearer's support system if assistance is needed. We collected physiological data from 107 veterans who volunteered to participate during five Project Hero bicycle riding events across the U.S. between August 2017 and April 2018. This analysis included data from 100 participants due to data corruption in the remaining seven participants' recordings. Participants ranged from 24 to 74 years of age (M = 47.3, SD = 11.0). Events took between three and seven days, during which participants were asked to wear a smartwatch (Moto 360 Sport or Polar M600) equipped with the tool for the duration of the event. The tool was designed to operate and receive user feedback continuously. Participants received training at the beginning of events and were asked to report the occurrence of PTSD triggers by tapping anywhere on the watch face. The tool recorded this ground truth as time-stamped events for further analysis.

In addition to the self-reported PTSD trigger events, our dataset also contains continuous time-series accelerometer and heart rate data. Accelerometer data were sampled at a frequency of 60 Hz; heart rate data were collected at 10 Hz. Despite the requirements for continuous usage of the device, data were not recorded at all times due to issues with the device batteries. The length of the recordings for each of the participants varied, with an average length of 14.58 hours (SD = 15.01 hours). As this work represents an initial exploration of the use of smartwatch-based physiological measures to predict PTSD triggers, we focused on only the heart rate data for this analysis. Analysis of the demographic factors and accelerometer data is left for future work.

2.2. Data preprocessing

All data preprocessing was completed in R version 3.5.1 (2017). The data preprocessing consisted of four steps: windowing, window labeling, Kalman filter imputation, and division of the data into training and testing datasets. These steps are discussed in the remainder of this section.

2.2.1. Windowing

In order to provide sufficient patterns of data for stress moment classification, the heart rate data were divided into one-minute sliding windows with 50% overlap. This method is a standard approach for time-series data (Dietterich, 2002). This window size was selected based on prior analyses of heart rate data (McDuff *et al.*, 2014) and the maximum resolution of data imputation (discussed in the following section). The overlap was selected based on prior analyses (Salahuddin *et al.*, 2007), and was intended to reduce the correlation between features in successive windows. Each window was assigned a label based on the presence or absence of a self-reported PTSD trigger event within the window. Windows containing a PTSD trigger event were assigned a PTSD trigger label, and windows without an event were assigned a normal label. After windowing, the dataset consisted of 11,312 windows with 172 PTSD event windows.

2.2.2. Kalman filter imputation

One of the challenges with off-the-shelf wearable technology is the presence of missing data. We addressed this issue

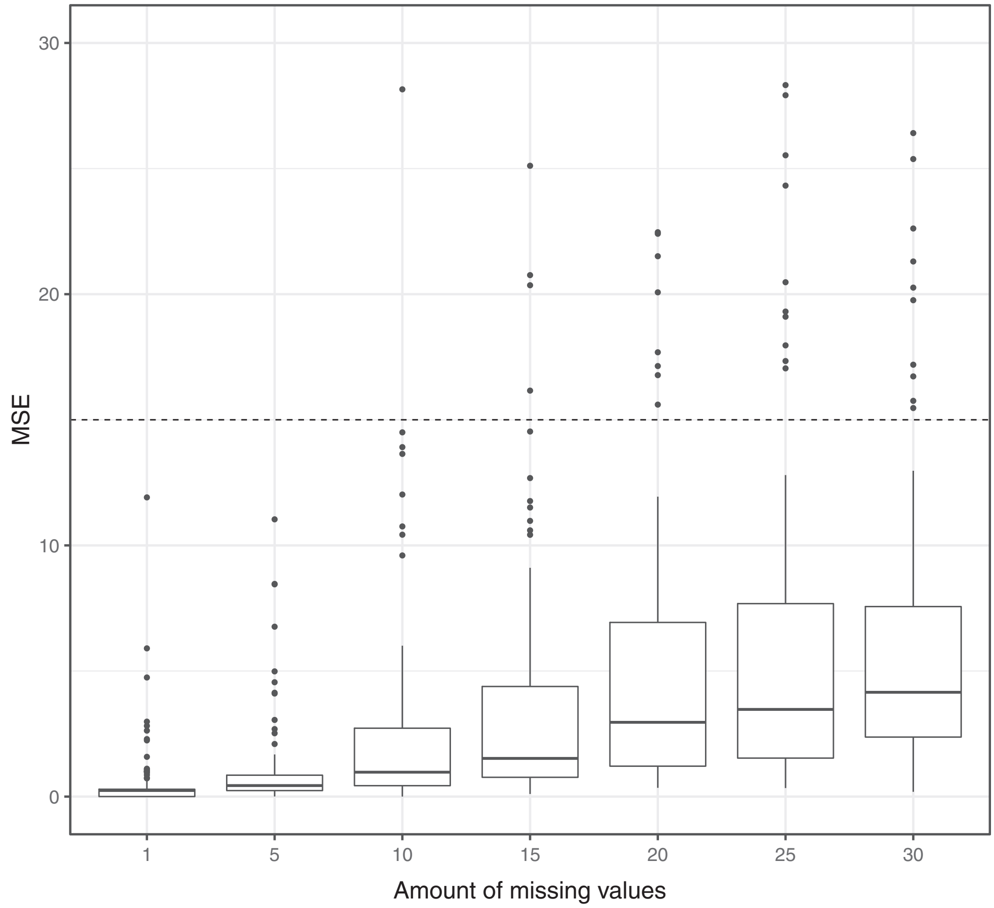

Figure 1. Mean Squared Error (MSE) for Kalman filter projections of known data with ranges of consecutive missing values. The horizontal dashed line represents the maximum MSE cutoff (15).

Table 2. PTSD trigger events and non-PTSD trigger events in the training and testing datasets.

Label	Train	Test
Non-PTSD Triggers	610	133
PTSD Triggers	122	50

with Kalman filter imputation, which is based on quadratic dynamic systems (Chatfield, 2016). The imputation was completed with the "ImputeTS" package in R (Moritz and Bartz-Beielstein, 2017). While there are several methods for data imputation, Kalman filters were selected as they have been validated for HR imputation in previous work (Gui *et al.*, 2014). We further evaluated this decision with an analysis of imputation results and known data.

The known data imputation analysis was conducted with complete one-minute windows. Ranges of consecutive HR values were randomly dropped in these windows and the dropped values were projected using Kalman filtering. Following the imputation, the average mean squared error between the imputed data and original data was calculated. We investigated ranges of consecutive missing values from 1 to 30. The results of this analysis are shown in Fig. 1. Based on this analysis, we selected 5 as the maximum consecutive imputation range, as it was the largest set of consecutive values to have a maximum MSE under 15, a cutoff used in Gui *et al.* (2014). Following this step, all windows with more than five consecutive missing values were dropped from the dataset. The final number of complete windows in the dataset was 10,081 with 172 PTSD event windows.

2.2.3. Training and test sets

In order to assess algorithm effectiveness, the dataset was randomly divided into training and testing sets by participant. Separating at the participant level avoids bias in the algorithm toward unique artifacts in participants' heart rate profiles and gives the best possible estimate of how the results would generalize to a broader veteran population. A split of 70% training and 30% testing was used to ensure that enough PTSD stress events remained in the testing set for adequate statistical power. A final downsampling step was performed on the training data to rebalance the amount of positive and negative training instances to a ratio of 5-to-1. This step was used to reduce bias and improve efficiency in the algorithm training process. The choice of a 5-to-1 ratio was assessed with a sensitivity analysis that compared model performance for models trained on a 5-to-1, 2-to-1, and 1-to-1 ratio of PTSD stress events and non-PTSD stress events. This analysis indicated that a 5-to-1 ratio led to the best predictive performance across algorithms. The numbers of data instances in training and testing are shown in Table 2.

Table 3 Summary of the features included in the non-neural network based algorithms.

Feature number	Feature name	Feature description
1	FFT coefficient 0	The real component of the absolute value of the first coefficient of a discrete Fourier decomposition of the signal
2	FFT coefficient 19	The real component of the absolute value of the nineteenth coefficient of a discrete Fourier decomposition of the signal
3	FFT coefficient 26	The real component of the absolute value of the twenty-sixth coefficient of a discrete Fourier decomposition of the signal
4	FFT coefficient 28	The real component of the absolute value of the twenty-eighth coefficient of a discrete Fourier decomposition of the signal
5	FFT aggregated variance	The variance of the absolute Fourier transform spectrum.
6	FFT aggregated skew	Skew of the absolute Fourier transform spectrum.
7	Energy ratio	The sum of squares of chunk 2 out of 10 chunks of the time-series expressed as a ratio with the sum of squares over the whole series
8	Change quantiles	The average of consecutive changes of the heart rate time-series inside of a corridor between quantiles 0.4 and 0.6 of the distribution of heart rates in the window.
9	Aggregated linear trend	The linear least-squares regression for values of the time series that were aggregated over chunks versus the sequence from 0 up to 50.

2.3. Feature generation and selection

Previous research has indicated that statistical features (Boonnithi *et al*., 2011), spectral features (Zheng *et al*., 2016), and nonlinear features (Melillo *et al*., 2011) are all valid approaches for feature generation in this context. We used a comprehensive time-series feature generation and selection process in this study that leveraged the TSFRESH package in Python (Christ *et al*., 2016). The TSFRESH package generates a feature set including distributional parameters (e.g., mean, quartile range), Fourier and wavelet components, and other features, such as sample entropy, the time of the last maximum value, and the frequency of the minimum and maximum values. The feature set that TSFRESH calculates includes all of the feature types discussed in Table 1. The feature generation process was applied to the windowed Kalman filtered heart rate data for both the training and testing datasets. Following feature generation, a subset of the most relevant features was identified using the Benjamini Hochberg procedure (Benjamini and Yekutieli, 2001)—a hypothesis testing-based method that identifies the most relevant features in a large feature set. The Benjamini Hochberg process uses a null hypothesis that a feature has no predictive power and conducts a statistical test to attempt to reject the null hypothesis. Features with a p-value of less than 0.05 are retained in the final feature set. The testing data were withheld from the Benjamini Hochberg procedure to avoid bias in the final prediction results. After applying Benjamini Hochberg procedure-based feature reduction, the feature set included nine features—six representing Fourier components and one each representing aggregated linear trend, change quantiles, and energy ratio. Table 3 provides a description of these features. This feature data set was used in all algorithm fitting processes except for the neural network models, which used raw data following the example in Acharya *et al.* (2017b). The feature selection process was evaluated by training models with both the full feature set and the reduced feature set. In all cases, the reduced feature set improved algorithm performance.

2.4. Algorithm optimization and fitting

The literature on stress detection algorithms suggests that SVM, decision trees, and neural networks are promising approaches. One goal of the present analysis is to assess the performance of CNN and random forests, especially as they compare to less complex algorithms (i.e., neural networks and decision trees). Another goal is to assess how the techniques employed here compare to known benchmarks. SVMs provide a benchmark as they have been used extensively in similar comparative analyses (e.g., McDuff *et al*., 2014; Zhai and Barreto, 2006). Therefore, we included five algorithms in this analysis: SVM with a radial kernel, decision trees, neural networks, CNN, and random forests. The SVM, decision tree, and random forest models were fit with the feature data using the caret package in R (Kuhn *et al*., 2017). The algorithm fits were optimized via an internal cross-validation using the training data that used AUC as the objective criteria. The CNN and neural network algorithms were fit to the raw data using the Python package Keras (Chollet, 2015). The architecture and parameters for the CNN were optimized to maximize AUC on a validation set using the Python package Hyperas (Pumperla, 2018). The algorithms and their optimized parameters and algorithms are summarized in Table 4. All algorithms were designed to provide a continuous prediction of PTSD stress moment likelihood as output. While the structural interpretation of most of the algorithms is clear from the information in Table 4, the CNN is less apparent. Figure 2 depicts the architecture of the CNN along with the dimensions of each layer. The layers are identified by their layer type: convolutional, max pooling, flattening, dense (or fully connected), and output. The convolutional layers filter the input data and transform it to a new feature space. The max pooling layers reduce the dimensionality of these new feature spaces to expedite learning. The flattening layer converts the pooled features into a two-dimensional array so that they can be processed with the dense layers. The dense layers and output layer form a standard multilayer neural network, which converts the new feature space into predictions.

3. Results and discussion

The algorithms were evaluated by their Area Under the receiver operating characteristic Curve (AUC). The receiver operating characteristic (ROC) curve is a plot of true positive rate by false positive rate over a range of algorithm confidence thresholds. The AUC is a robust measure of binary classification performance, as it is insensitive to class disparities (Fawcett, 2004). Statistical differences in the ROC

Table 4 Fitted algorithms and optimized parameters.

Machine learning approach	Algorithm input	Parameters optimized	Parameter values
SVM	Heart-rate features	Cost (C)	1
		Sigma	10
		Kernel	Radial Basis Function
Decision tree (DT)	Heart-rate features	Maximum tree depth	19
Neural network	Kalman-filtered heart rate data	Number of hidden layers	2
		Neurons in each hidden layer	HL1 = 32, HL2 = 18
		Weight decay	0.01
		Batch size	64
		Epochs	2000
		Activation function	Sigmoid
Random forest	Heart-rate features	Size of feature set	6
CNN	Kalman-filtered heart rate data	Architecture	See Figure 2
		Weight decay	0.01
		Batch size	6
		Epochs	2000
		Activation function	Sigmoid

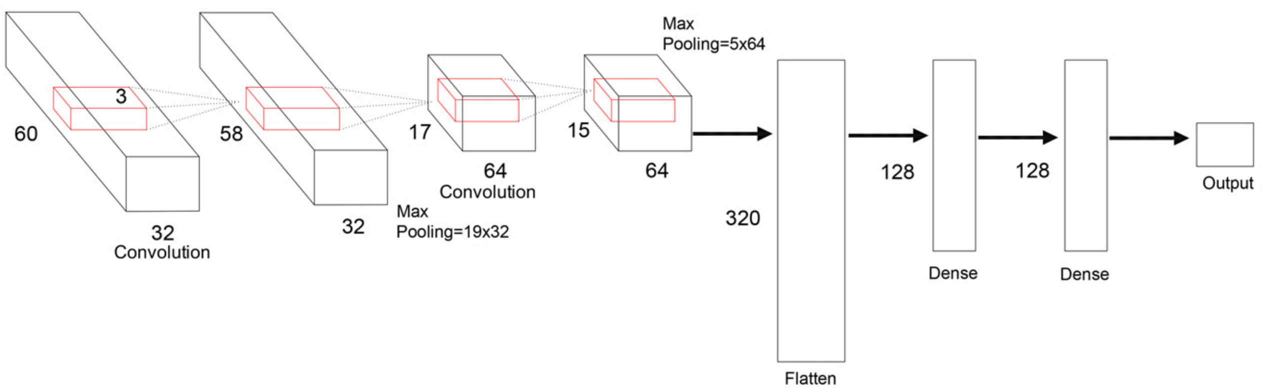

Figure 2. CNN architecture diagram indicating the order and dimensions of the convolutional, max pooling, and fully connected layers.

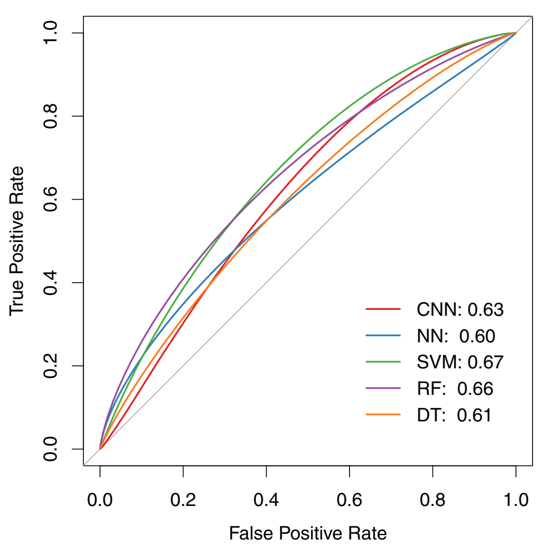

Figure 3. ROC curves and corresponding AUC values for the five algorithms evaluated in this study.

curves were assessed using a one-sided DeLong's test (DeLong *et al.*, 1988), conducted with the "pROC" package in R (Robin *et al.*, 2011). A threshold of $p = 0.05$ was used for all tests.

3.1. Algorithm predictive performance

Figure 3 shows the ROC curves and AUC values for the five fitted algorithms. All algorithms predicted significantly better than random. The SVM with the radial basis kernel had the highest AUC, 0.67, although the random forest was within 0.01 AUC, and the CNN was within 0.04 AUC. Pairwise comparisons indicated no significant differences between the algorithms.

Collectively, these results provide support for the use and further development of SVM, ensemble techniques (such as random forests), and deep learning models (such as CNN) for PTSD symptom detection. These results align with previous explorations of SVM and deep learning and stress detection (McDuff *et al.*, 2014; Singh *et al.*, 2013); however, these findings are unique in focusing on PTSD and in comparing other common machine learning algorithms.

3.2. Prediction analysis

One limitation of the deep learning and ensemble approaches is that they provide limited insight into the patterns of heart rate associated with PTSD trigger onset. In order to assess such differences qualitatively, we analyzed the predictions on the raw heart rate data from the CNN, SVM, and random forest models. These results are shown in Fig. 4, which plots raw, mean-centered, heart rate traces for correctly predicted windows with a summary trendline (in blue). The traces for the PTSD symptom data are centered on the onset of symptoms (30 s into the window). Figure 4 shows that, despite significant noise in the data, the algorithms appear to associate an increase in heart rate with

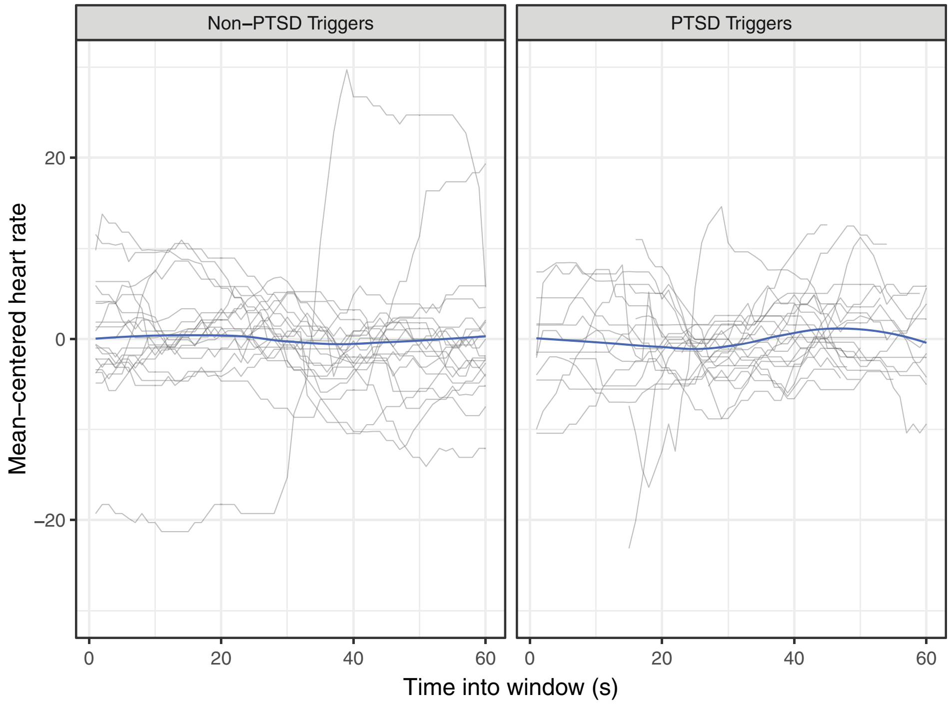

Figure 4. Heart rate profiles for correctly predicted PTSD symptoms and non-PTSD symptom moments. Predictions are based on the CNN algorithm.

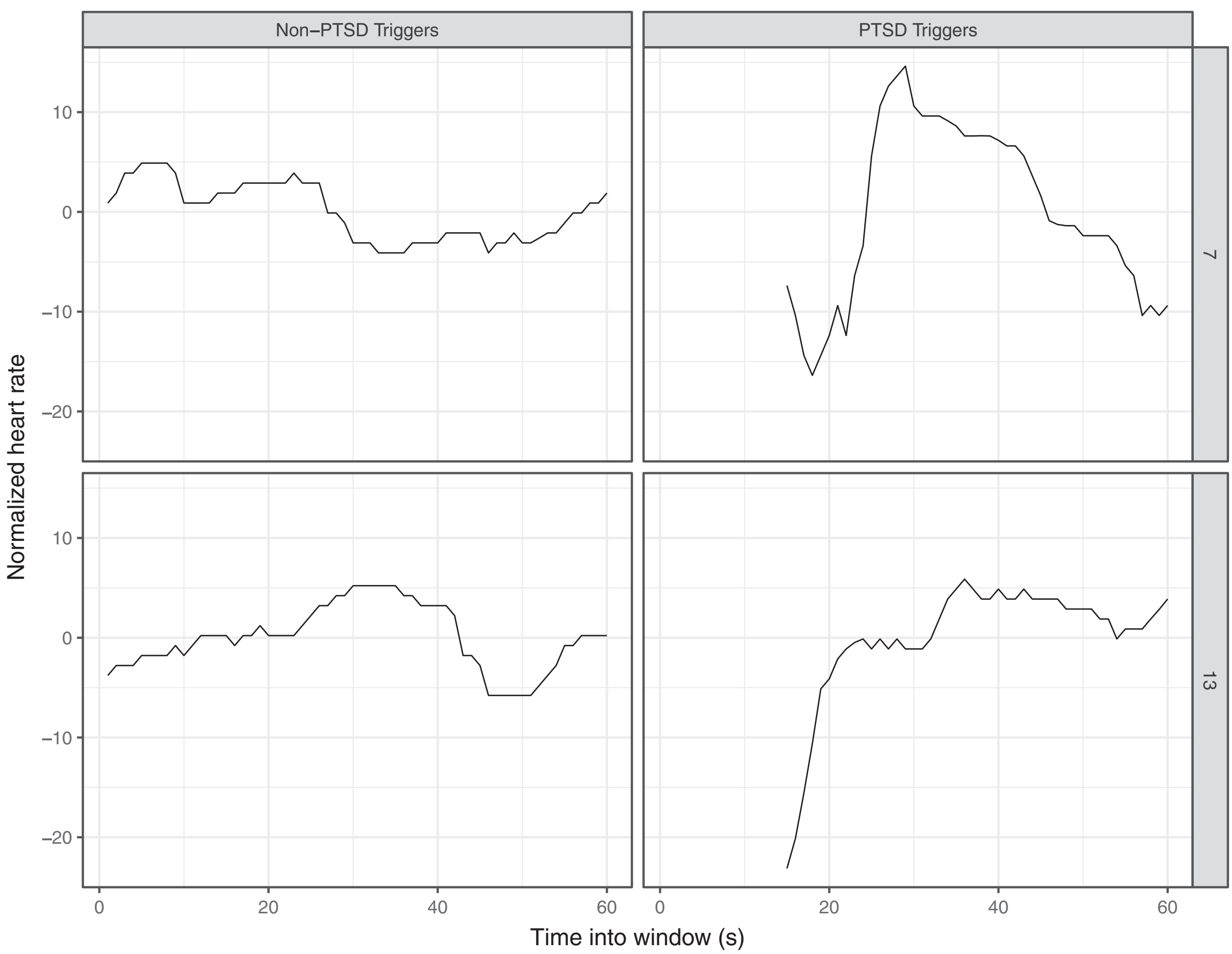

Figure 5. Exemplar windows (window 7 and window 13) of non-PTSD trigger events and PTSD trigger events. In the PTSD windows, the event occurs at a time into the window of 30 s.

PTSD trigger onset. These trends are further highlighted in Fig. 5, which shows exemplar windows of data containing PTSD trigger events and non-PTSD trigger events. This finding aligns with reviews of PTSD symptom onset, which suggest heart rate accelerations following the onset of PTSD symptoms (Khanade and Sasangohar, 2017).

3.3. Limitations and future work

The approach documented in this article represents our initial computational work for detecting PTSD triggers from physiological data in a naturalistic setting. While the results show promise in real-time detection of PTSD triggers, the approach is limited by the dataset as well as the scope of approaches explored here. Most importantly, while participants were briefed to report PTSD trigger events, the sensitivity to, and interpretation and tolerance of, such events may vary among patients, resulting in nonhomogeneous data. Such variability is expected at the current preliminary stage of research; however, future work should aim at personalized deep learning solutions that take individual characteristics into consideration. In addition to reporting limitations, the continuous operation of the tool proves taxing on Android devices' battery life. In fact, some participants reported battery life as low as four hours. Given the busy schedule of activities during the riding events, recharging the tool might have been impractical and challenging. In addition, a significant amount of data was missed due to perspiration (e.g., during the rides), damage to the devices (e.g., due to rain), misuse (e.g., not wearing the device properly), forgetting to wear the device (e.g., after charging the watch or showering) or not wearing the device at all. Another potential threat to internal validity of the ground truth data collection method used is overreporting. While the participants were asked to only report the occurrence of PTSD triggers, some participants might have also reported general high anxiety/stress feelings. In addition, while all participants claimed being diagnosed with PTSD, the research team had no viable method for verification of such claims.

From an analysis perspective, this work may be limited by the imputation methods, downsampling, and reliance on heart rate data. Future work should explore methods such as Gaussian Process Regression—as assessed in Gui *et al.* (2014)—for data imputation. These methods may yield a larger dataset, which, in turn, could reduce the likelihood of overfitting. Similarly, alternative sampling techniques, such as Synthetic Minority Over-sampling (Chawla *et al.*, 2000) or weighted samples, could be used to create larger training sets and further reduce the likelihood of overfitting. Future work should analyze additional measures such as acceleration and heart rate variability, following the example of Sano and Picard (2013). Acceleration data may be used as a corollary to physical activity and disambiguate heart rate fluctuations caused by activity rather than PTSD symptoms.

Despite these important limitations, to the best of our knowledge, the technology that uses the algorithm documented in this article is the first attempt at non-invasive continuous monitoring of PTSD triggers. While other tools and methods exist to monitor general stress (as discussed in the introduction), such tools are not specific to PTSD, do not use deep learning or ensemble techniques, and generally suffer from intrusiveness, are non-discreet, and in most cases lack evidence suggesting that a user-centered design, development, and analysis method is employed. Work is in progress to collect data to build on the promising results documented in this article to improve the sensitivity and specificity of the detection algorithm, as well as to investigate the acceptance of such technology among veterans.

4. Conclusions

PTSD is a prevalent condition among combat veterans and has resulted in significant societal challenges in the United States. While tools and technologies exist to help veterans manage their symptoms, continuous monitoring tools to detect the occurrence of PTSD triggers and to intervene in a timely manner are missing, despite potential value. Our overall research aims to address this important gap by utilizing non-intrusive sensing, user-centered and persuasive design, and machine learning methods to develop an "always-on" intervention to support veterans suffering from PTSD. To this end, a novel tool has been designed to collect ground truth from veterans via naturalistic testing. This article documented the development and analysis of series of algorithms that used heart rate collected from this tool to predict the onset of PTSD symptoms. The results suggest that SVM, random forests, and CNN algorithms predict PTSD symptom onset significantly better than random classifiers. Analysis of the predictions suggests that the algorithms associated increases in heart rate with PTSD trigger onset. While several limitations remain and need to be addressed in future work, the promising results presented in this article align with previous analyses of stress detection from physiological data, suggesting that the approach may be generalized to broader contexts involving stress detection and mitigation.

Acknowledgment

The authors would like to thank Carolina Rodriguez-Paras, Kunal Khanade, Eric Czarnecki, Patricio Rodriguez-Paras, and Abby Hutton for their help during data collection and data clean-up.

Funding

This research was partly funded by a Texas A&M University Engineering Experiment Station (TEES) Interdisciplinary Research Seed Grant awarded to Dr. Farzan Sasangohar.

ORCID

Anthony D. McDonald http://orcid.org/0000-0001-7827-8828
Farzan Sasangohar http://orcid.org/0000-0001-9962-5470
Ashish Jatav http://orcid.org/0000-0001-8549-2224
Arjun H. Rao http://orcid.org/0000-0002-6682-6812

References

Acharya, U. R., Fujita, H., Lih, O. S., Adam, M., Tan, J. H., and Chua, C. K. (2017a) Automated detection of coronary artery disease using different durations of ECG segments with convolutional neural network. *Knowledge-Based Systems*, **132**, 62–71.

Acharya, U. R., Fujita, H., Oh, S. L., Hagiwara, Y., Tan, J. H., and Adam, M. (2017b) Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals. *Information Sciences*, **415–416**, 190–198.

Acharya, U. R., Lih, S., Hagiwara, Y., Hong, J., and Adeli, H. (2017c) Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. *Computers in Biology and Medicine*, **100**, 1–9.

Al Ayubi, S. U., Parmanto, B., Branch, R., and Ding, D. (2014) A persuasive and social mHealth application for physical activity: A usability and feasibility study. *JMIR mHealth uHealth*, **2**, e25. doi: 10.2196/mhealth.2902

American Psychiatric Association. (2013) *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. Washington, DC: American Psychiatric Pub.

Baker, J. E. (1984) Monitoring of suicidal behavior among patients in the VA health care system. *Psychiatric Annals*, **14**, 272–275.

Bauman, W., and Melnyk, W.T. (1994) A controlled comparison of eye movements and finger tapping in the treatment of test anxiety. *J. Behav. Ther. Exp. Psychiatry*, **25**, 29–33.

Benjamini, Y., and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**(4), 1165–1188.

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

Bousefsaf, F., Maaoui, C., and Pruski, A. (2013) Remote assessment of the heart rate variability to detect mental stress. *IEEE*. doi:10.4108/icst.pervasivehealth.2013.252181

Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 5–32.

Chatfield, C. (2016) *The Analysis of Time Series: An Introduction* (6th ed.). CRC Press, Boca Raton, FL.

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

Chollet, F. O. (2015). *Keras: A Deep learning library for theano and tensorflow*. https://github.com/fchollet/keras. (accessed June 30, 2018).

Christ, M., Kempa-Liehr, A. W., and Feindt, M. (2016) Distributed and parallel time series feature extraction for industrial big data applications. http://arxiv.org/abs/1610.07717

Ciabattoni, L., Ferracuti, F., Longhi, S., Pepa, L., Romeo, L., and Verdini, F. (2017) Real-time mental stress detection based on smartwatch. *2017 IEEE International Conference on Consumer Electronics* (ICCE), Las Vegas, NV, 8–10 January 2017, 110–111. doi:10.1109/ICCE.2017.7889247

Costin, H., Rotariu, C., and Păsărică, A. (2012) Mental stress detection using heart rate variability and morphologic variability of ECG signals. *2012 International Conference and Exposition on Electrical and Power Engineering (EPE 2012)*, 591–596.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, **44**, 837–845.

Dietterich, T. G. (2002) Machine learning for sequential data: A review. In *Structural, Syntactic, and Statistical Pattern Recognition*, 15–30. Springer, Berlin, Germany.

Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S., and Acharya, U. R. (2018) Deep learning for healthcare applications based on physiological signals: A review. *Computer Methods and Programs in Biomedicine*, **161**, 1–13.

Fawcett, T. (2004) ROC graphs: Notes and practical consideration for researchers. *ReCALL*, **31**, 1–38.

Foa, E. B., Chrestman, K. R., and Gilboa-Schechtman, E. (2008) *Prolonged Exposure Therapy for Adolescents with PTSD Emotional Processing of Traumatic Experiences, Therapist Guide*. Oxford University Press, New York, NY.

Forbes, D., Creamer, M., Bisson, J. I., Cohen, J. A., Crow, B. E., Foa, E. B., … Ursano, R. J. (2010) A guide to guidelines for the treatment of PTSD and related conditions. *Journal of Traumatic Stress*, **23**, 537–552.

Gui, Q., Jin, Z., and Xu, W. (2014) Exploring missing data prediction in medical monitoring: A performance analysis approach. In *2014 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 1–6.

Healey, J. A., and Picard, R. W. (2005) Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, **6**, 156–166.

Hendin, H. and Haas, A. P. (1991) Suicide and guilt as manifestations of PTSD in Vietnam combat veterans. *The American Journal of Psychiatry*, **148**, 586.

Holmgard, C., Yannakakis, G. N., Karstoft, K.-I., and Andersen, H. S. (2013) Stress detection for PTSD via the StartleMart Game. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 523–528.

Khanade, K. and Sasangohar, F. (2017) Efficacy of using heart rate measurements as an indicator to monitor anxiety disorders: A scoping literature review. *Proceedings of the Human Factors and Ergonomics Society: Annual Meeting Human Factors and Ergonomics Society Meeting*, **61**, 1783–1787.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., … Hunt, T. (2018) Package "caret" classification and regression training. https://cran.r-project.org/web/packages/caret/caret.pdf. (accessed June 30, 2018).

Lawrence, C. E., Reilly, A. A., Quickenton, P., Greenwald, P., Page, W. F., and Kuntz, A. J. (1985) Mortality patterns of New York State Vietnam veterans. *American Journal of Public Health*, **75**, 277–279.

LeCun, Y., and Bengio, Y. (1995) Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks*, **3361**, 255–258.

McDonald, A. D., Lee, J. D., Aksan, N. S., Dawson, J. D., Tippin, J., and Rizzo, M. (2017) Using kinematic driving data to detect sleep apnea treatment adherence. *Journal of Intelligent Transportation Systems*, **21**, 422–434.

McDonald, A. D., Lee, J. D., Schwarz, C., and Brown, T. L. (2014) Steering in a random forest. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **56**, 986–998.

McDuff, D., Gontarek, S., and Picard, R. (2014) Remote measurement of cognitive stress via heart rate variability. *Proceedings of 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2957–2960.

Melillo, P., Bracale, M., and Pecchia, L. (2011) Nonlinear heart rate variability features for real-life stress detection: Case study: Students. *Biomedical Engineering Online*, **10**, 96.

Melillo, P., Formisano, C., Bracale, U., and Pecchia, L. (2013) Classification tree for real-life stress detection using linear heart rate variability analysis: Case study: Students under stress due to university examination. *World Congress on Medical Physics and Biomedical Engineering*, Beijing, China, 477–480.

Moritz, S., and Bartz-Beielstein, T. (2017) imputeTS: Time series missing value imputation in R. *The R Journal*, **9**, 207–218.

Picard, R. W., Vyzas, E., and Healey, J. (2001) Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**, 1175–1191.

Pumperla, M. (2018) Hyperas. https://github.com/maxpumperla/hyperas

R core team. (2017) R: A language and environment for statistical computing (Version R version 3.4.1 (2017-06-30)). R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org

Rand, D., Eng, J. J., Tang, P. F., Jeng, J. S., and Hung, C. (2009) How active are people with stroke? Use of accelerometers to assess physical activity. *Stroke*, **40**(1), 163–168.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011) pROC: An open-source package for R and S + to analyze and compare ROC curves. *BMC Bioinformatics*, **12**, 77. doi: 10.1186/1471-2105-12-77

Rodriguez-Paras, C., Tippey, K., Brown, E., Sasangohar, F., Creech, S., Kum, H.-C., … Benzer, J. K. (2017) Posttraumatic stress disorder and mobile health: App investigation and scoping literature review. *JMIR mHealth and uHealth*, **5**, e156.

Salahuddin, L., Cho, J., Jeong, M. G., and Kim, D. (2007) Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings. *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Lyon, France, 22–26 Aug. 2007, 4656–4659.

Sano, A., and Picard, R. W. (2013) Stress recognition using wearable sensors and mobile phones. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, Geneva, Switzerland, 2–5 Sept. 2013, 671–676.

Singh, R. R., Conjeti, S., and Banerjee, R. (2013) A comparative evaluation of neural network classifiers for stress level analysis of automotive drivers using physiological signals. *Biomedical Signal Processing and Control*, **8**, 740–754.

Sloan, D. M., Gallagher, M. W., Feinstein, B. A., Lee, D. J., and Pruneau, G. M. (2011) Efficacy of telehealth treatments for posttraumatic stress-related symptoms: A meta-analysis. *Cognitive Behaviour Therapy*, **40**, 111–125.

U.S. Department of Veterans Affairs. (2016) How common is PTSD? National Center for PTSD. https://www.ptsd.va.gov/public/PTSD-overview/basics/how-common-is-ptsd.asp (accessed August 21, 2018).

U.S. Department of Veterans Affairs. (2017) What is PTSD? National Center for PTSD. https://www.ptsd.va.gov/public/PTSD-overview/basics/what-is-ptsd.asp (accessed August 21, 2018).

Williams, J. P. C., Moon, J., Sasangohar, F., Khanade, K., and Rodriguez-Paras, C. (2018) Veteran-centered design of a post-traumatic stress disorder (PTSD) tool: Qualitative analysis of interviews with veterans. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, **62**(1), 548–549.

Zanon, M., Sparacino, G., Facchinetti, A., Riz, M., Talary, M. S., Suri, R. E., … and Cobelli, C. (2012) Non-invasive continuous glucose monitoring: Improved accuracy of point and trend estimates of the multisensor system. *Medical & Biological Engineering & Computing*, **50**(10), 1047–1057. doi:10.1007/s11517-012-0932-6

Zhai, J., and Barreto, A. (2006) Stress detection in computer users through non-invasive monitoring of physiological signals. *Biomedical Sciences Instrumentation*, **42**, 495–500.

Zheng, Y., Wong, T. C. H., Leung, B. H. K., and Poon, C. C. Y. (2016) Unobtrusive and multimodal wearable sensing to quantify anxiety. *IEEE Sensors Journal*, **16**, 3689–3696.