IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 January 2023, accepted 23 February 2023, date of publication 8 March 2023, date of current version 16 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3254134

RESEARCH ARTICLE

Real-Time Personalized Physiologically Based Stress Detection for Hazardous Operations

TOR T. FINSETH[1,5,6], (Member, IEEE), MICHAEL C. DORNEICH[2,5,6], (Senior Member, IEEE), STEPHEN VARDEMAN[2], NIR KEREN[3,5,6], AND WARREN D. FRANKE[4,5,6]

[1]Department of Aerospace Engineering, Iowa State University, Ames, IA 50011, USA
[2]Department of Industrial Manufacturing and Systems Engineering, Iowa State University, Ames, IA 50011, USA
[3]Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, IA 50011, USA
[4]Department of Kinesiology, Iowa State University, Ames, IA 50011, USA
[5]Human Computer Interaction, Iowa State University, Ames, IA 50011, USA
[6]Virtual Reality Application Center, Iowa State University, Ames, IA 50011, USA

Corresponding author: Michael C. Dorneich (dorneich@iastate.edu)

This work was supported in part by the U.S. National Aeronautics and Space Administration under Grant 80NSSC18K1572.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Iowa State Institutional Review Board.

ABSTRACT When training for hazardous operations, real-time stress detection is an asset for optimizing task performance and reducing stress. Stress detection systems train a machine-learning model with physiological signals to classify stress levels of unseen data. Unfortunately, individual differences and the time-series nature of physiological signals limit the effectiveness of generalized models and hinder both post-hoc stress detection and real-time monitoring. This study evaluated a personalized stress detection system that selects a personalized subset of features for model training. The system was evaluated post-hoc for real-time deployment. Further, traditional classifiers were assessed for error caused by indirect approximations against a benchmark, optimal probability classifier (Approximate Bayes; ABayes). Healthy participants completed a task with three levels of stressors (low, medium, high), either a complex task in virtual reality (responding to spaceflight emergency fires, n =27) or a simple laboratory-based task (N-back, n =14). Heart rate, blood pressure, electrodermal activity, and respiration were assessed. Personalized features and window sizes were compared. Classification performance was compared for ABayes, support vector machine, decision tree, and random forest. The results demonstrate that a personalized model with time series intervals can classify three stress levels with higher accuracy than a generalized model. However, cross-validation and holdout performance varied for traditional classifiers vs. ABayes, suggesting error from indirect approximations. The selected features changed with window size and tasks, but found blood pressure was most prominent. The capability to account for individual difference is an advantage of personalized models and will likely have a growing presence in future detection systems.

INDEX TERMS Stress detection, machine learning, physiological sensors, virtual reality, spaceflight training.

I. INTRODUCTION

Despite extensive training in responding to an emergency, a person's response to an actual emergency can be negatively affected by the stressfulness of the situation. Stress can result in a cascade of physiological changes that may alter behavioral patterns, situational awareness, decision making, and cognitive resources [1]. An inability to cope with the stress of a high-stress condition can decrease task performance and thereby risk mission failure, injury, or death [2]. Consequently, developing resiliency to this situational stress through improved training may lead to better outcomes. To that end, using real-time monitoring of a person's stress responses to customize the stressfulness of training scenarios

The associate editor coordinating the review of this manuscript and approving it for publication was Filbert Juwono.

may, in turn, lead to more appropriate handling of actual hazardous operation [3], [4].

Stress detection using machine learning has been challenging for several reasons. First, there are individual differences in the appraisal of, and physiological responses to, stressful situations. Numerous stress detection approaches have attempted to reduce technical complexity by generalizing their models to a broad population, or the ''average'' response [3]. However, the stress response to a unique situation is largely subjective, and personalized stress detection models may be more robust to individual differences [5], [6].

The second challenge is that the time series nature of physiological signals can be problematic. The physiological stress response has temporal and feature correlations. These correlations may violate the machine learning assumption that the data are independently and identically distributed, thereby leading to biased results [7].

An additional challenge is interpreting how well model estimations match the true conditional probabilities of a subject's stress levels. Stress detection models rely on traditional machine learning algorithms that make data-driven approximations to estimate the chance that the individual is experiencing a state of stress given their physiological responses. However, these estimations are often indirect and without a benchmark for comparison. From classical statistics research, the Bayes theorem is theoretically the optimal solution and a classifier given the same parameters as Bayes theorem will have the lowest probability of error [8]. The Bayes theorem uses an empirical density distribution as a true prior probability, which can be used to calculate the conditional probability of each class. The classifier selects the class with the greatest posterior probability of occurrence, also known as maximum a posteriori. Machine-learning algorithms attempt to approximate the density distributions. If the density estimates of the classifier converge to the true densities, then the estimated probability represents the true probability of occurrence and a classifier that approximates Bayes becomes an Optimal Bayes classifier. However, these approximations can have varying accuracy due to assumptions made by the algorithm, such as independence of predictors [9]. Thus, it can be difficult to interpret the model's logic. Physiological systems are known to have a high degree of dependence with regard to a stress response, because they are often initiated by the same neuroendocrine axis [10]. Some researchers have shown that classifiers may account for dependencies using multivariate kernel density estimators [11]. Therefore, it may be beneficial to evaluate supervised machine learning classifiers against a benchmark optimal classifier that approximates Bayes using a density distribution estimated through multivariate kernel density estimation for stress detection.

To achieve real-time and continuous monitoring of stress levels, new approaches are needed to analyze time series for physiologically-based stress detection [12]. Real-time stress detection can enable closed-loop automation to either modify the training environments to better match the trainee's responses or better assess individual stress during staged or real operations [13]. In datasets with repeated measurements at multiple times that present uncertainty from randomness or incompleteness, such as multiple measures of physiological data, multivariate kernel density estimators may help increase detection accuracy [11].

To address these challenges, the goal of this research is to assess the objectivity, reliability, and validity of a personalized model methodology. The first research question focuses on objectivity, and whether the stressor levels can show distinct levels in personalized features used for the classification model while accounting for individual differences in physiology. This will provide confidence that the model is designed for the appropriate context and that the training data reflect distinct ground truth levels. The second research question focuses on the system's reliability by evaluating the performance of the time-series interval approach using a post-hoc model comparing between a standard laboratory cognitive task and a complex job-specific task, window sizes, classifier validation techniques, and features selected for each individual. The third research question focuses on the validity of the system by seeking to understand whether indirect approximations influence traditional supervised machine learning classifiers compared to a Bayes classifier, known as Approximate Bayes (ABayes), which uses direct approximations of optimal stress classes through multivariate kernel density estimation.

This research is part of a larger development effort to design VR training scenarios that can dynamically adapt a virtual environment using real-time stress detection [14], [15], [16]. To answer these research questions within the constraints of the larger system, the experiment will assess a time-series interval approach to stress detection for a post-hoc model of physiological response data, its accuracy in detecting participant stress using a collected during stressful tasks, and provide the architecture for a real-time stress detection system that uses this classification methodology. Validating a machine learning pipeline post-hoc allows for translation to real-time stress detection and applications for stress monitoring.

II. BACKGROUND

Stress detection systems rely on classifying physiological signals into multiple stress classes using machine learning. However, stress detection is challenging due to the individual differences in the response to stress and the time-series nature of the physiological stress response. This section describes the physiological responses, stress detection research and physiological sensors, approaches to classifying time series data, a deeper look at the challenges facing stress detection, and the current research approach.

A. TIME COURSE OF THE STRESS RESPONSE

The physiological stress response involves the interaction between the nervous system and the endocrine system that aims to maintain physiological integrity under changing environmental demands. The time course of the physiologic

responses to stress varies by system and by the intensity and duration of the stressor; they are neither physiologically independent nor statistically orthogonal. After the psychological appraisal of a stressor, neural ganglia pathways are activated almost instantaneously to evoke very rapid responses via local neurotransmitters. For example, disinhibition of heart rate via vagal withdrawal occurs within milliseconds while a sympathetically-mediated increase in heart occurs after a few seconds (5-10 s) [10]. Sympathetic and sudomotor activity results in the opening of eccrine sweat glands on hands and feet, which occur about 1-5 seconds after stimuli [17]. On the other hand, the physiologic responses due to circulating chemicals take longer to manifest. Epinephrine is secreted from the adrenal medulla and range from milliseconds to minutes to exert their cardiovascular effects. Whereas, cortisol is initiated by the adrenal cortex 5–10 min after stressor onset and peak between 20 and 30 min [18]. These processes can act exclusively or in conjunction on target organs to potentiate (e.g., memory, muscle activation) or attenuate organ function (e.g., digestion, reproduction).

There is increasing support that the physiological systems activated are those best suited to cope with the type of stressor, rather than the prior theories that certain systems are activated if the stressor magnitude surpasses a threshold [19], [20]. For example, stress caused by a traffic jam may cause one individual to show an increase in epinephrine while another individual may show an increase in the stress hormone cortisol with little to no increase of epinephrine. Therefore, the same stressor can differentially affect individuals via leading to the activation of varying physiological systems, with each system having personalized, and differing, times-scales to respond and recovery. The different individual stress response and system time-scales present challenges in detecting and classifying levels of stress.

B. STRESS DETECTION

Stress detection, by means of classifying these physiological responses into levels of stress via machine learning, continues to evolve and is motivated by the potential utility of continuously monitoring stress levels in real-time [12], [21]. Stress detection systems have been developed for drivers in semi-urban scenarios [22], [23], patients undergoing virtual reality therapy [24], individuals in working environments [25], and people that need help managing daily stress [21], [26], [27], [28], [29], [30]. Stress detection can also be applied to a variety of human-machine interfaces (HMIs) which may monitor stress, but also infer the cognitive state of the user to adapt system functionality [31]. Examples of HMIs that may use stress detection include wearable devices, voice recognition systems, eye tracking systems, facial expression analysis, and brain/body computer interfaces [12], [32]. However, these HMIs may not be able to accurately detect stress in all individuals, and the accuracy of stress detection may vary depending on the specific technology and approach used [33].

These detection systems collect information about stress responses from either objective physiological sensors or subjective psychological metrics, in the form of independent variables called features, which are then used to classify the stress level. Commonly used sensors include electrodermal activity (EDA), electrocardiogram (ECG), respiration (RSP), electroencephalogram (EEG), skin temperature (ST), and blood volume pulse (BVP) [33]. For an ECG signal, stress indices have been primarily inferred from changes in the time intervals between heartbeats, which measure Heart Rate Variability (HRV) using time-domain, frequency-domain, or nonlinear analysis. HRV metrics have been associated with sympathetic and parasympathetic activation. However, attempting to detect stress levels from signal amplitude alone neglects the time series nature of physiological data. Physiological systems may be simultaneous and coupled (e.g., breathing can modulate heart rate), contain both deterministic and stochastic components, and may be correlated when measured over long periods of time [34]. Stress sensor signals are continuous ordered attributes; therefore, they are best characterized by features that quantify the distribution of data points, variation, correlation properties, stationarity, entropy, and nonlinear properties [35].

C. APPROACHES TO TIME SERIES CLASSIFICATION

To address the time series nature of physiological signals, common time series classification methods include a) comparing whole series data by employing distance-based algorithms like Dynamic Time Warping (DTW), b) performing high-level feature extraction from successive, sequential time intervals and classifying intervals with a model, c) judging the presence or absence of short patterns (i.e., shapelets) in the whole series, d) frequency counts of recurring patterns to form a ''dictionary'' that defines the classes, e) combinations of the aforementioned methods, or f) model-based learning methods like those relying on auto-regressive models or hidden Markov models [36]. Each of these methods has advantages and disadvantages with respect to physiological stress detection.

DTW is highly effective with a nearest-neighbor classifier for time series data, such as repeated patterns in ECG due to heart arrhythmias and sleep apnea. However, there are few examples of its application to stress detection [37]. This is likely due to acute stress having temporal and pattern variation, which make it difficult for whole series, shapelets, or dictionary methods to be effective. Model-based learning methods, like hidden Markov models, fit multiple models to the data in order to determine the best model to use. This type of framework has been seldom studied for physiological stress detection unless paired with stress speech analysis [30]. Interval characteristics is the most common classification method for stress detection. It is implemented by extracting features for windows/epochs, which is a highly reliable analytical method for quantifying stress through features like HRV [38], [39]. Typically, the time window size is

predetermined. However, some features may behave differently depending on the window size. For example, HRV frequency features are recommended to have windows in the order of minutes and smaller time intervals may increase error [39]. An evaluation of window sizes can help identify which features work best for interval methods.

Neural networks have become a popular classifier choice for interval methods, due to highly accurate frameworks such as convolutional or recurrent neural networks [40]. The success of a neural net is partly due to its ability to handle unequal time series lengths and optimize model parameters over time [41]. However, neural networks simultaneously extract features, and many of the classification rules are created by the model rather than by programmers. These classification rules can be hidden within interconnected layers [12]. The net effect is that the logic used in the classifications is often implicit and uninterpretable. For this reason, traditional machine learning models that classify interval features from a time-series are more informative and interpretable as to how data points are assigned to classes. Interval classification often uses supervised learning, where classification models are trained using interval features separated into classes/states (e.g., low, medium, high stress levels) and the model is subsequently used to detect class labels based on class/state probability of a test dataset. Traditional supervised machine learning algorithms include support vector machine (SVM), decision tree, and random forest.

D. CHALLENGES OF PHYSIOLOGICAL STRESS CLASSIFICATION

A major challenge in using physiological signals for detection is the rigidness of generalized models in accounting for physiological differences between people. Stress varies among individuals due to differences in appraisals of the stressor and the perceived threat, but also the body's capability to enact the physiological responses. For example, an EDA-based generalized classifier that is deployed and tested on multiple people may have higher classification error among a subset of this group, since as much as 25% of the population are EDA non-responders or hypo-responders [42]. By not accounting for differences in physiology, inherent errors are created when using generalized models for physiological detection. This challenge has led some researchers to believe that personalized models may be more accurate [3], [12]. Revising the example, higher accuracy may be achieved by the EDA-based classifier if the model accounts for the individual's respective EDA level and reactivity, or instead rely on other sensors when EDA is not a reliable predictor for that individual. While EDA is one of many physiological systems, some may be more susceptible to individual differences than others (e.g., cortisol [43]). Supervised classifiers can be personalized by having the stress detection system create a model using training data from the individual and by selecting discriminate and relevant features for the individual [44].

Another challenge is that supervised classifiers have a degree of uncertainty depending on how they estimate probability distributions in order to label stress levels. Supervised models produce a probability distribution for each stress level (class) for a set of physiological signal data points (vectors); this distribution determines which class is most probable at a given time. However, rather than creating a distribution directly from the dataset, the probability distribution is created indirectly (and often ad hoc) based on the technical specifics of a classification method. For example, decision tree classifiers produce rectangles that partition the input space and calculate the approximate class probabilities based on the number of vectors located within each rectangle. Thus, the class probability is constant for each rectangle and always discontinuous at the rectangle boundaries, leading to a probability that is more defined by how the rectangles are positioned within the input-space rather than the vector distribution across the entire input-space. Similarly, SVMs create a hyper-planes intended to produce maximum separation between class vectors in the input space. Ad hoc ''approximate class probabilities'' are often created using softmax functions of distances from vectors to hyperplanes—a practice that may not match empirical probability estimates [45]. The process by which these ad hoc methods approximate class probabilities does not easily translate to meaningful cause/effect insights related to either changes in the environment or the measured changes in physiological measurements.

The translation of a post-hoc system (i.e., offline) to real-time (i.e., online) brings another set of challenges commonly associated with data collection in ambulatory settings that are less controlled. One major challenge is the need to process and analyze data in real-time, which requires a system with high computational power and efficient algorithms that have minimal loss of data and error propagation during data analysis. Another challenge is the need to transmit data from the sensors to the system in real-time, which requires a reliable and high-speed wireless network [12]. Ensuring the privacy and security of the data is another important consideration, as the data may contain sensitive personal information and could be vulnerable to cyber-attacks. Additionally, there may be challenges in accounting for environmental context, as the physiological indicators of stress may be affected by other factors such as physical activity, medication, and ambient temperature.

Any classifier can be used with a personalized detection approach, but the classifier selected should maximize the confidence that the approximate class probabilities match empirical probability estimates. Since Bayes theorem provides more direct estimations of conditional probabilities, its effects are more interpretable and may provide insight into whether the aforementioned traditional classifiers have error resulting from indirect approximation. This can be achieved by implementing the Bayes theorem in a new approximately Bayes classifier (ABayes). To that end, along with a real-time personalized stress detection system, the secondary goal of this research is to assess the extent to which traditional supervised machine learning methods (decision tree, support

vector machine, and random forest classifiers) are limited compared to an optimal probability; a classifier based on Bayes theorem using multivariate kernel density estimates.

E. APPROACH

This paper describes the development of a personalized physiological-based stress detection system to classify acute stress using feature selection on intervals of the time-series data. To train the machine learning model, participant physiological signals were collected for three stressor levels during either a spaceflight emergency fire procedure on a VR International Space Station (VR-ISS) [46], [47] or a well-validated and less-complex N-back mental workload task [48]. Several previous studies have detected stress induced by N-back tasks via machine learning methods, both alone [48], [50] and with another job-specific task [51]. Therefore, comparing a job-specific VR-ISS task to the N-back using the same personalized approach is a way to assess the system's reliability can work for multiple stress detection tasks. Each participant had features selected at different interval window sizes, then those personalized features trained the classifier model, and subsequently tested the classifier's predictive accuracy. Since the stress response is complex and often unique, the analysis will explore which features are selected most for individuals depending on window size, and how this changes classification performance. Classifier performance was assessed using both holdout and cross-validation validation techniques to simulate how the model may perform on unseen data as an analog for deployment in real-time.

The novelty and contribution of this research is to show that stress detection may benefit from using personalized time-series approaches to quantify temporal patterns in physiological signals, to assess whether traditional classifiers are limited in approximating the optimal Bayes solution, that certain features may be better at different windows sizes, and that this approach has a suitable performance for detecting stress for a VR spaceflight emergency training procedure.

III. METHODS

A. PARTICIPANTS

Forty-one healthy participants (34 male, 7 female) performed a complex task in virtual reality (spaceflight emergency fire, N = 27) or a laboratory-based task (N-back, N = 14). The mean age was 20.9±6.5 years, all adults in the age range of 18-41 years. The demographic distribution included 76% European American/White, 12% Asian or Asian American, and 7% Hispanic or Latino. All study procedures were approved by the Institutional Review Board of Iowa State University.

B. EXPERIMENTAL DESIGN

The evaluation had two types of tasks and three stressor levels for each task. Task was a between-subjects variable: participants conducted either a fire response task aboard a VR International Space Station (VR-ISS) or a computer-based N-back task. These tasks were selected since it is possible to facilitate varying degrees of task complexity. Stressor level was a within-subjects variable: each task consisted of three stressor levels (low, medium, and high). Trials were counterbalanced via Latin Squares, which uses a grid of numbers or letters representing different conditions in the study to assign participants equally and prevent trial order effects [52].

FIGURE 1. VR-ISS emergency fire (low, medium, high stressor scenarios).

One task, VR-ISS, is the virtual reality environment of the ISS specifically designed for participants to implement an emergency fire response procedure by locating and extinguishing a fire source [46]. The VR-ISS task is highly dynamic and in a complex environment with many stimuli and task steps. The task is based on existing NASA Emergency Procedures [53] but simplified to reduce the amount of needed training. Several dynamic interactions were included in the VR-ISS to aid detection and location of the fire source. To locate the fire, participants evaluated atmospheric contaminant levels, which changed as a function of time and distance from the fire source. The highest contaminant value indicates the approximate location of the fire source. Thus, participants would have to monitor and recall local contaminant levels. When needed, participants used virtual oxygen masks and fire extinguishers.

Stressor levels in the VR-ISS were created using a combination of environmental stressor intensities that were independent of the task procedure: smoke, alarm noise, and flickering module lights [47]. The low stressor level did not contain any stressors; therefore, a voice recording announced a fire situation at the beginning of the simulation. The medium stressor level included a continuous caution alarm, low smoke density (visibility limit of 6 ft.) and flashing lights in one of the three ISS modules. The high stressor level involved a continuous caution alarm, a continuous fire alarm, flickering lights in all modules, and dense smoke (visibility limit of 1 ft.). Fig. 1 presents the smoke density in the VR-ISS for each stressor level. Prior research verified that the three stressor levels produced different levels of subjective stress [47].

The other task, N-back, is presented with a sequence of colored squares on a computer screen; participants need to recall the location of the square that was shown n steps earlier in the sequence. The N-back task is a well-validated stressor [48] where low-complexity can be induced through manipulating the one primary stressor of working memory demand. Stress is manipulated by asking participants to recall 1-back (low-demand), 2-back (medium-demand), and 4-back (high-demand). The N-back task is a measure of working

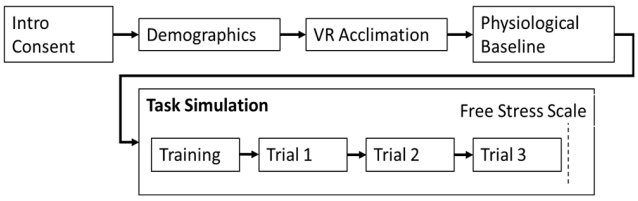


FIGURE 2. Design and procedure of the study.

memory capacity that is associated with executive functioning which can affect physiological stress indices [53].

C. STRESS MANIPULATION MEASURE

To verify that the three stressor levels produced variable levels of stress, the participants completed the Free Stress scale after the third trial. This scale was used to rate the subjective stress level on a scale of 0 to 100 (least to most stressful) [23], [54], [55]. The stress appraisal process is continuous and relative, with reappraisals of the experience happening long after the stressful exposure [56]. The Free Stress scale was intended to relatively measure the subjective stress by comparing all three trials at the same time.

D. PROCEDURE

The experiment was completed in a single laboratory visit, lasting approximately 120 minutes (Fig. 2). After participants provided written consent, they completed a demographic questionnaire. To acclimate to VR before the data collection tasks, participants were placed in the VIVE Virtual Reality Home simulation [57] to train them to navigate, operate, and control the VR simulation. Participants were asked about their cybersickness. Participants were equipped with physiological sensors and a baseline recording was taken to verify that the sensors were working properly.

For the task simulation (Fig. 2), participants were then assigned to either the VR-ISS or N-back tasks. For the VR-ISS, participants completed a 20-30 minute VR tutorial that included information about the VR-ISS layout, how to navigate, fire equipment, and the emergency fire response. For N-back, participants completed a 3 minute tutorial on how to indicate whether the current stimulus was the same as the one presented N trials ago. Participants then completed three trials: low, medium, and high stressor levels. Participants were given 5 minutes between trials to recover to physiological baseline. The Free Stress scale was administered after the last trial.

E. OVERVIEW OF THE STRESS DETECTION SYSTEM

To aid in the development of the stress detection system, a machine learning pipeline was developed to detect and classify the three stress levels from the physiological measures and to evaluate ABayes as a classifier against other supervised classifiers. The pipeline consisted of several steps including data collection, preprocessing, feature extraction, feature selection, and classification as presented in Fig. 3. Data were collected through multiple sensors that measure physiological responses. A time-series classification approach was implemented by segmenting the data into multiple intervals and using summary measures as features. A feature extraction process was then used to find a high-level subset of features that may have class discrimination with respect to a single individual, which was reduced into a low-dimensional feature subset (feature selection) by means of classification of a random holdout. A supervised approach was then taken to train the classifiers with the selected feature subset comprised of physiological data from three stress trials for each participant, investigated only as a subject-specific personalized model. Lastly, the classifiers were evaluated on their ability to detect participant stress levels.

The modification of an offline detection system to an online system may have a simple goal of outputting stress predictions in real-time. However, stress predictions may also inform adaptive systems about adjust automation to optimize human-computer performance, such as to enhance the user knowledge retention during task training or lessen a pilot's cognitive workload while flying a plane [16]. As an applied example, the machine learning pipeline used for offline validation was modified to collect and classify stress level in real-time (Fig. 4). Post-hoc classification was modified to be real-time continuous classification. The individual's training model (i.e., data from selected features during post-hoc validation; Fig. 3) is imported. This dataset includes the participant's physiological signals in each of the three stress classes, all collected prior to using the real-time system to predict stress. The signal preprocessing for the online real-time system used the exact same algorithms as the signal preprocessing of the offline system for a 30-second window, thereby ensuring the validity of the feature calculations. The real-time stress system uses a parallel processing architecture to first configure equipment sequentially, then to execute processes simultaneously. To configure the equipment, worker 1 configured the Biopac MP150 system (Biopac Systems Inc., Santa Barbara, CA) parameters for each sensor, worker 3 configured the TCP connections between clients and server and imported the selected features and trained model, then worker 2 configured the AcqKnowledge software (Version 5.0.1, Biopac Systems Inc.) to receive these signals and commanded the Biopac to begin streaming data. Once the command was sent to stream data from the Biopac, the data were continuously buffered and classified into real-time windows. The window classification labels were then used for post-classification logic (e.g., automation adaptations) and output (i.e., human-computer interface). Since this system was intended for acute stress detection (i.e., $<$ a few hours), the stream data were only classified using an initial model, and retraining was not incorporated into the system design.

The online real-time system was run with a separate participant sample within a larger adaptive VR stress training system (study results are reported in [15]). The adaptive system was designed to detect stress and then provide immediate adaptive feedback during training to increase/decrease

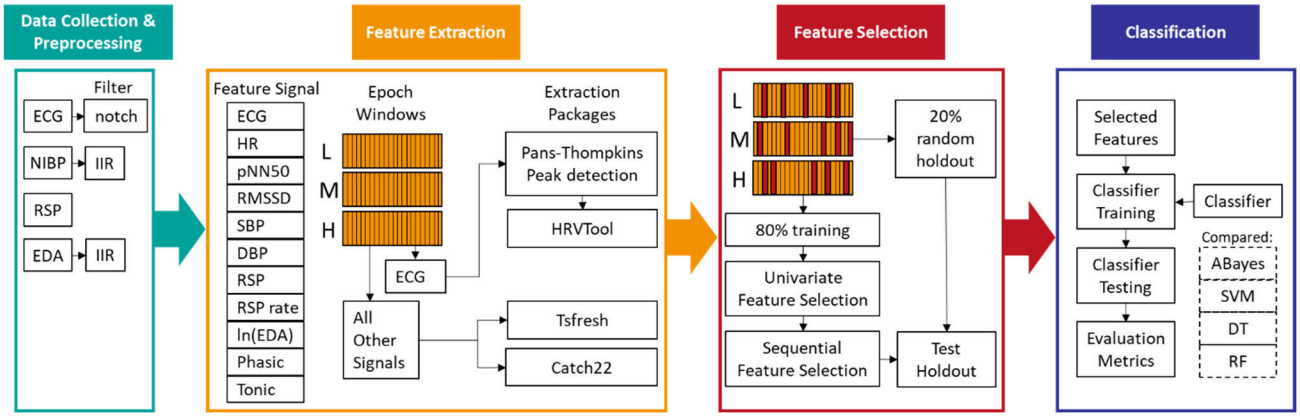


FIGURE 3. The Post-hoc machine learning pipeline of stress detection and classification.

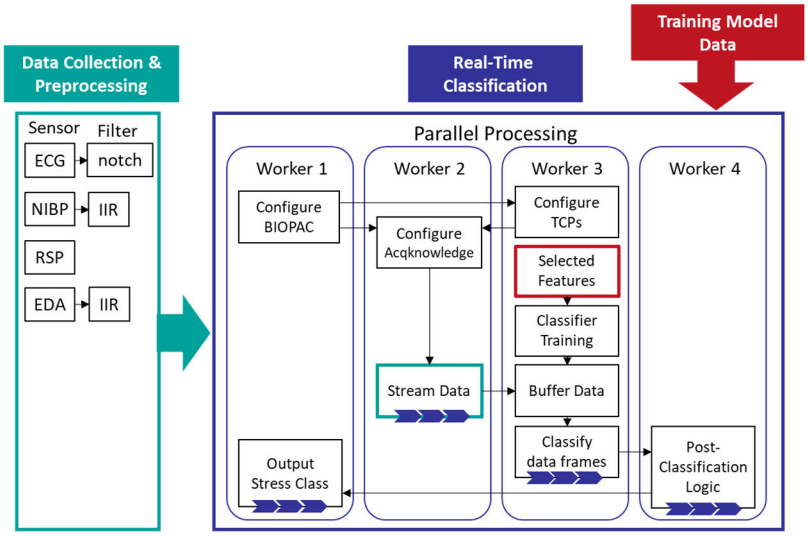


FIGURE 4. Stress detection implemented as a real-time pipeline [14], [15]. Blue arrows represent processes running in parallel that the continuous data streams. All other processes within the workers are sequential. Colored boxes represent inputs from other components.

the stressors in virtual reality to help build competency and inoculate stress. Since the ground truth of the adaptive system constantly changed due to training effects, the offline evaluation of holdout and cross-validation was used to better assess the reliability and validity of the stress detection described in this paper.

To test the validity and reliability of the online system, the feature selection and classification accuracy were compared to those of the offline system, and a latency test was performed. Using an offline model and predetermined individuals' stress data, both offline and online systems produced the same selected features and accuracy. This result was to be expected as the only programmatic difference for the online system is how the computation is distributed across parallel workers. A second test was performed to determine the computational latency of online system workers. The classification time for all classifier algorithms was less than the signal buffering timeframe (e.g., 30-seconds), ensuring that the online system did not propagate latency errors when retrieving a consecutive buffer window for classification.

F. DATA COLLECTION & PREPROCESSING

Data were collected for the machine-learning pipeline using four physiological signals that were acquired simultaneously: electrocardiogram (ECG), Electrodermal Activity (EDA), Respiration (RSP), and Noninvasive Blood Pressure (NIBP). Biopac MP150 system (Biopac Systems Inc., Santa Barbara, CA) was used to measure ECG, and was equipped with an ECG100C module [58]. ECG and RSP were sampled using Biopac MP150 (125 Hz) and Bionomadix Bioshirt that uses Bluetooth signal thereby increasing mobility of the participant. Beat-to-beat blood pressure data were collected using an oscillometric noninvasive blood pressure (NIBP) fingercuff placed on the participants' nondominant hand over the middle phalanx of the long and ring finger (CNAP Monitor 500, CNSystems Medizintechnik AG). The nondominant arm was

placed in an arm sling to standardize the position of the hand relative to the heart between all participants. To calibrate the finger cuff, an NIBP cuff (CNAP Monitor 500, CNSystems Medizintechnik AG) was placed on the participant's dominant upper arm and measured periodically to minimize potential hydrostatic pressure differences between the fingers and heart level. NIBP was sampled with the Biopac MP150 at 125 Hz. Electrodermal activity (EDA) measures changes in electrical conductivity in the skin due to production of sweat by activation of the autonomic nervous system (ANS). Increased arousal during stress will elicit higher EDA. Electrodes were placed on the intermediate phalanges on the index and middle fingers of the nondominant hand. EDA was sampled with the Biopac MP150 (125 Hz).

As expected, the data contained different types of noise and artifacts associated with subject movement, power line, and electromagnetic interference. NIBP was corrected for motion artifacts using an IIR band pass filter with cut-off frequencies at 1 Hz and 10 Hz. The ECG signal was filtered for electrical noise by an internal 50-60 Hz notch filter. The EDA signal was corrected with an IIR low pass 2nd order Butterworth filter fixed at 5 Hz [59], [60], [61].

G. FEATURE EXTRACTION

The feature extraction process was intended to improve information density by extracting a variety of features that characterize the time-series data. The feature extraction used by this study involved two steps: (1) derive feature signals that are physiologically relevant to the stress response and then (2) extract generalized features chosen with the intent to characterize distribution of data points, variation, correlation properties, stationarity, entropy, and nonlinear properties [35].

First, the raw signals were used to extract feature signals with AcqKnowledge software (Version 5.0.1, Biopac Systems Inc.). The ECG signal was used to calculate Heart Rate and two time-domain Heart Rate Variability (HRV) signals of root mean squared of successive differences (RMSSD) and percent of peak-to-peak intervals exceeding 50 milliseconds (pNN50). Increasing values of RMSSD and pNN50 indicate relaxation (vagal activation) and decreasing values indicate arousal (vagal inhibition). RMSSD and pNN50 features were extracted from the ECG signal. Respiration was also measured as an indicator of ANS activity. Systolic blood pressure (SBP) and diastolic blood pressure (DBP) were extracted from NIBP as another measure of cardiovascular reactivity. DBP and SBP can reflect changes in the total peripheral resistance of blood vessels. Increases in local sympathetic activity cause constriction of blood vessels, while reductions in sympathetic activity lead to dilation. In the absence of changes in cardiac output, decreases in blood vessel constriction are usually reflected by decreases in DBP. EDA can be parsed into slower tonic-level and faster changing phasic-level components. Skin Conductance Level (SCL) is a measure of tonic EDA and reflects the general changes in autonomic activity. Skin Conductance Response (SCR) is discrete, short, phasic fluctuations that reflect higher frequency variability of the signal as a response to immediate stimuli [62]. The tonic component was extracted by low pass filtering with a cut-off frequency of 0.16 Hz, while the phasic component was extracted with a band-pass filter of 0.16 Hz and 2.1 Hz [53]. A smoothing window of 5-second averages was used on all derived feature signals (e.g., HR, RMSSD, SBP, DBP) to reduce the potential adverse impact of possible outliers from momentary sensor disconnecting from the skin, electromagnetic interference, and algorithm detection error.

The raw sensor signals and feature signals were saved at 125Hz and binned into epoch windows from which other signals were extracted. Since the goal of this study was to build an automatic stress classification model with the potential to be applied to real-time applications, small window sizes were selected for evaluation: 10 sec, 20 sec, 30 sec, and 40 sec.

The second step of the features extraction process is to extract features from the time series. Two toolset packages were used: Tsfresh [63] and Catch22 [64] for automatic feature extraction of time series characteristics, including absolute energy, absolute sum of changes, autocorrelation, entropy, and number of values above and below the average. Since ABayes is designed for probability density estimation in a real-time system, Tsfresh features were excluded if they were Boolean data types or were previously reported to take longer than 10-2 second to compute (see [63]). The Catch22 toolset is a set of 22 time series features from the much larger MATLAB toolbox, called *hctsa*, which has high accuracy in detecting different types of time series data [64]. The features extracted by Catch22 and Tsfresh do not overlap. From the ECG signal, the inter-beat interval (RR) signal was extracted via Pans-Tompkins peak detection [65]. Time-series and spectral HRV features were then extracted from the RR signal via HRVTool [66]. The Pans-Tompkins MATLAB code and HRVTool outputs were checked and found to be reliable compared to the peak-detection and HRV calculation in AcqKnowledge and Kubios HRV MATLAB toolbox (Version 2.2) [67]. The final extracted features are listed in the Appendix.

H. FEATURE SELECTION

Feature selection is the process of reducing the dimensionality of the classification problem by finding an optimal subset of available features that provide class discrimination. The best subset contains the fewest number of dimensions that most contribute to the classifier performance; the remaining, less contributing features are discarded [68].

A hybrid method of feature selection was conducted with a two-step process involving a univariate feature selection (UFS) filter method and sequential feature selection (SFS) wrapper method together in series [44]. Feature selection was only implemented on a training dataset (as opposed to testing dataset) to mitigate data leakage where the model may be over-fit and overestimate the model performance when

deployed [69] (see Data Analysis). To improve the robustness of the feature selection, a portion of the training dataset was held out for validating the selection process.

For the wrapper holdout, 20% of epochs were randomly selected and stratified from each class. These were set aside to ensure that a test set of equal class sizes remained unseen while the other 80% was used to select features. Feature selection was conducted on the remaining 80% using the combination of UFS and SFS. First, UFS was used to find the best features for classification by quantifying their discriminative power using a univariate statistical test. The features were then ranked according to their mean one-way ANOVA F-value to prioritize features that explain large amounts of variance. In the second step, the SFS method starts by iteratively adding features from the UFS in a forward search to measure performance gain. SFS starts with the most discriminate feature identified by UFS and then adds features, one-by-one, according to their F-value rank and stopped after 12 iterations. When a feature is added to the subset, 10-fold cross-validation was performed, where misclassification rate was used as a criterion to opt for the best subset of features. The classifier used in the SFS was the same as the respective classifier used in the final validation techniques.

Since the wrapper holdout was selected randomly, running the feature selection multiple times would produce different optimal feature subsets. Therefore, this entire process of UFS and SFS was repeated six times and the final features were those that appeared in the six optimal features subsets more than twice. Upon conclusion of the feature selection, the wrapper holdout was again included in the dataset to prepare for classification training and testing.

I. CLASSIFICATION

The ABayes classifier was formulated to assess the indirect ways that standard machine learning algorithms typically estimate probability distributions across classes for given input vectors. When considering traditional machine learning classifiers, all standard classifier development and performance evaluation is implicitly or explicitly done using a probability model that generates class-conditional probabilities. To create a distribution of probability densities, the probability model states that for classes $k = 1, 2, \ldots, K$, observable data vectors $\boldsymbol{x}$ are generated for each random choice of a class $y = k$ by using probability distributions. The class probability distribution is specified by a priori class probabilities $\pi_1, \pi_2, \ldots, \pi_K$. Then, data vectors $\boldsymbol{x}$ are generated for a given class via a class-conditional probability density $g_k^{i_k}(x_i)$. For a classifier $f(\boldsymbol{x})$ that maps observations to classes $(f(\boldsymbol{x}) \in \{1, 2, \ldots K\}\ \forall \boldsymbol{x})$, the conditional probabilities can be calculated as in (1).

$$P[f(\boldsymbol{x}) = y] = \sum_{k=1}^{K} \pi_k \int_{f(\boldsymbol{x})=y} g_k(\boldsymbol{x})\, d\boldsymbol{x} \tag{1}$$

In an optimal situation, the densities $g_k(\boldsymbol{x})$ and probabilities π_k are perfectly known to the classifier. Therefore, state k is classified with maximum $\pi_k g_k(\boldsymbol{x})$ (2), which is equivalent to the maximum conditional probability (i.e., maximum a posteriori) of state/class k given the observation $\boldsymbol{x}$ (4), which is also equivalent to the maximum a posteriori of the optimal classifier (4). An optimal classifier with perfectly known a priori probabilities and density distributions is also known as Bayes Optimal Classifier, in which the accuracy would be higher than all other approximations [9].

$$\underset{k}{\operatorname{argmax}}\, \pi_k g_k(\boldsymbol{x}) \tag{2}$$

$$\underset{k}{\operatorname{argmax}} \left(\frac{\pi_k g_k(\boldsymbol{x})}{\sum_{l=1}^{K} \pi_l g_l(\boldsymbol{x})} \right) \tag{3}$$

$$\begin{aligned} f^{opt}(\boldsymbol{x}) &= \underset{k}{\operatorname{argmax}}\, P[k|\boldsymbol{x}] \\ &= \underset{k}{\operatorname{argmax}} \left(\frac{\pi_k g_k(\boldsymbol{x})}{\sum_{l=1}^{K} \pi_l g_l(\boldsymbol{x})} \right) \\ &= \underset{k}{\operatorname{argmax}}\, \pi_k g_k(\boldsymbol{x}) \end{aligned} \tag{4}$$

However, standard machine learning classifiers are not optimal and are limited because the densities $g_k(\boldsymbol{x})$ are not known. Subsequently, some classifiers attempt to make approximations of the post-data weights π_k and densities $g_k(\boldsymbol{x})$ for each state/class, while other classifiers refrain from estimating the distributions entirely and attempt to approximate $f^{\text{opt}}(\boldsymbol{x})$. For example, in tree algorithms, the relative frequencies of the training set class/state in rectangles serve as weight estimates of conditional probabilities of classes given that the input vector falls in given rectangles, whereas SVM directly learns a decision boundary without estimating data generating distributions. Even in Naive Bayes, the classifier estimates all marginal distributions and uses the product as a density, while making a generally poor assumption that the input vectors for each class have independent components [9]. In these cases, machine learning classifiers use data-derived functions of $\boldsymbol{x}$ to approximate density distributions as a substitute for the optimal classifier. Incidentally, vague density distributions can occur when the class relative proportions in training datasets are not same proportions as the π_k that exist in the optimal classifier. Probability densities can be adjusted so that training set class relative frequencies match desired weights (π_k) but parametrizing to class frequencies is typically unrealizable in practice. Therefore, no classifier can improve on this optimal classifier if the posterior weights and distribution (i.e., densities) are known.

The optimal classifier is derived from Bayes theorem, which provides a direct approximation of conditional class probabilities. Hence, the Approximate Bayes (ABayes) classifier is a statistical approach that attempts to optimally discriminate states/classes based on estimated conditional probabilities determined through a direct approximation for the multivariate kernel density estimates (Fig. 5). That is, training sets of multivariate observations from classes (k) had bandwidth (h) calculated and have been processed through a (Gaussian) kernel $(\mathcal{K})$ density estimation routine with features $DV = 1, 2, \ldots, m$, representing the multivariate

kernel to produce functions approximating the class conditional densities. A priori class probabilities approximated the optimal probabilities using class relative proportions in training datasets for experimental comparison to traditional machine learning algorithms. These are used to produce the Approximate Bayes classifier (Eq. 5), whereby an observation is classified to the class that gives it the largest estimated conditional class probability. Recognizing the limitations of other standard machine learning classifiers, Bayes should result in the most accurate probability estimate, all things being equal.

$$f^{opt}(\boldsymbol{x}) = \underset{k}{\operatorname{argmax}}\, \hat{\pi}_k \hat{g}_k(\boldsymbol{x}) \tag{5}$$

ABayes has fundamental differences with regard to well-known classifiers like Naïve Bayes. Naïve Bayes uses multivariate densities derived as products of marginal univariate density estimates (so that the multivariate densities are ones of independence for coordinates of their multivariate arguments), even when Naïve Bayes is based on kernel density estimation [70]. In comparison, ABayes creates multivariate density estimates (using product kernels) that will essentially never be of product form, because the product forms for the kernels do not force estimated densities to have product forms. This relies on the assumption that for training sets of n >1, the multivariate densities are not ones of independence for the coordinates of their multivariate arguments. Thus, the form of densities used in ABayes is discrete from Naïve Bayes classification.

The input variables of the classifier are the extracted features from physiological sensors, standardized in the range of {0, 1}. An estimate is made for the probability density per class by applying multivariate (Gaussian) kernel density estimates. The training phase initializes weights based on assumed frequency of occurrence. The test phase is used to automatically classify an unknown input vector.

J. GENERALIZED APPROACH

A generalized approach was also created to compare the performance of the personalized approach. To maximize the model performance, a subset of VR-ISS participants with similar data was selected (N = 7) using a participant-error curve. Features were predetermined by iteratively removing low accuracy features until an optimal feature subset remained (listed in the Appendix). Leave-one-subject-out validation was conducted on all participants and the results were averaged. The held-out test data were standardized using the mean and standard deviation of the training dataset.

K. DATA ANALYSIS

The stress detection capabilities of the novel classifier ABayes and three common supervised machine learning classifiers were compared: support vector machine (SVM), decision tree (DT), and random forest (RF). The classifiers were implemented with the MATLAB Statistics and Machine Learning Toolbox. The performance of the classifiers was evaluated using two different validation techniques: Cross-Validation and Holdout (Fig. 6). Cross-validation was performed by 10-Folds, with folds stratified and randomized. The holdout consisted of 20% of data from the end of each class being used as ''unseen'' testing data, which simulates how the system may perform when deployed. The cross-validation and holdout use only the participant's training dataset (i.e., training folds) for the feature selection. For example, only 90% of the data would be used for feature selection during 10Fold and then repeat with a new 90% partition for each fold, whereas the holdout technique selects features from the first 80% of each trial's dataset. The 10-fold cross-validation and holdout were implemented independently and use the hybrid combination of SFS and UFS on their respective training datasets and folds. Therefore, the feature selection was implemented after the data was partitioned for the chosen validation technique within the proposed machine learning pipeline (see Fig. 3). Further, the classifier chosen to be evaluated by cross-validation and holdout was the same classifier used for the feature selection, thereby mitigating wrapper bias [69]. All data was standardized prior to classification per the physiological signal per individual, with the testing data being standardized with respect to the means and standard deviations of the training data.

The evaluation process of classifiers involves calculating the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) [30]. Classifier performance was measured using accuracy, precision, recall, F1-score, and specificity. The metrics use multi-class classification with macro-averaging techniques [22].

Accuracy is one of the main performance indicators and is defined as the number of correctly classified labels divided by the total number of labels. The Precision, Sensitivity, and F1-score reflect the importance of the retrieval of positive labels, while the Specificity reflects the correct classification of negative labels. F1-score is regarded as a more reliable classifier performance metric in comparison to accuracy in some circumstances because the accuracy metric does not account for imbalanced class datasets [71].

The class size balance for each participant was evaluated with the imbalance ratio [72] and likelihood ratio imbalance degree (LRID) [73]. Imbalanced data can have harmful effects on classification and interpretation of results. The imbalance ratio (IR) is the most commonly adopted metric for class-imbalance extent, but it only for binomial datasets because it considers the ratio of the distribution of observations in the largest ($\hat{p}_{max}$) and smallest ($\hat{p}_{min}$) classes while ignoring information of other minority classes [72]. The frequency is estimated as the fraction of observations in a given class (n_k) divided by total number of observations (N). An IR of one suggests and equal dataset. LRID offers more resolution into multiple class distributions where the data may be overlapped or where there is ambiguity about the level of data separation. However, the score can vary by many magnitudes depending on the number of minority classes.

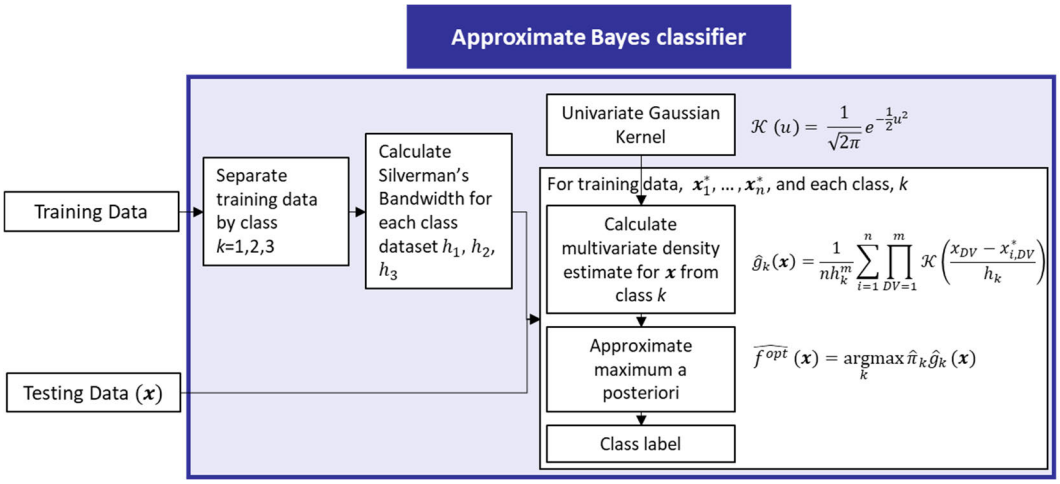


FIGURE 5. The approximate Bayes for stress level classification.

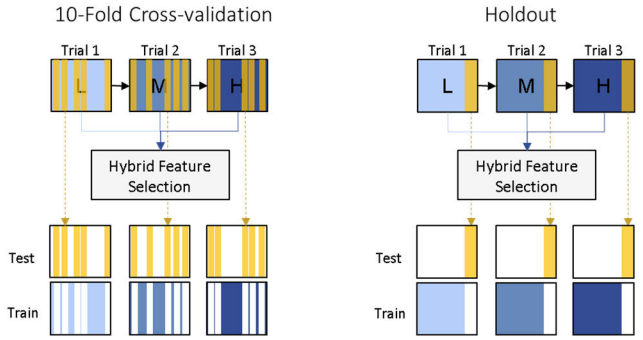


FIGURE 6. Examples of the feature extraction and selection for cross-validation and holdout. L, M, and H refer to the low, medium, and high stressor scenarios, respectively.

Likelihood ratios can range from zero to infinity, with a score of zero for balanced data, while imbalanced data will result in a score larger than zero. Likelihoods were converted to probabilities [74].

Data analysis on the subjective stress measure was performed using SPSS software (Version 23.0; IBM Corp.). Repeated measure analysis of variance (RM-ANOVA) was used to calculate the fixed effect of stressor level and pair wise comparisons that were adjusted to control for type I errors (Bonferroni adjustment). Results were considered significant for $p \leq 0.05$. Cohen's d was used for assessing effect size, where $0.2 < |d| < 0.5$ was considered a small effect size, medium effect size when $0.5 < |d| < 0.8$, and large effect size for $|d| > 0.8$ [75].

IV. RESULTS

A. SUBJECTIVE STRESS MANIPULATION

The main effect of stressor level on subjective stress was significant for the VR-ISS, $F(2,90) = 102$, $p < .001$, $d = 3.02$. All pairwise comparisons indicated the subjective stress was significantly different ($p < .001$) between the stressor levels. Similarly, the main effect of stressor level on subjective stress was significant for the N-back, $F(2,24) = 47.5$, $p < .001$, $d = 3.98$. Pairwise comparisons indicated subjective stress was significantly higher for participants in 4-Back compared to 1-Back ($p < .001$) and the 2-Back ($p < .001$). Subjective stress was significantly higher for 2-Back compared to 1-Back ($p = .018$).

B. MACHINE LEARNING RESULTS

The physiological data obtained from the VR-ISS and N-back were analyzed to provide insight into the features chosen by SFS, comparing the performance of ABayes between different tasks (VR-ISS, N-back), between different evaluation strategies (10-Fold, holdout), and compared to the three standard machine learning classifiers. The characteristics for each task dataset are listed in Table 1, including the number of 10-second windows averaged across participants. For the 20, 30, and 40-second windows, the windows were one half, one third, and one fourth of the number of 10-second windows.

When considering multiple classes, the LRID shows the VR-ISS is eleven times more likely to be imbalanced which is equivalent to 46% probability of being imbalanced. The N-back was only 1.65 times more likely to be imbalanced, which is 9.5% probability. Due to the probability of imbalance, the F1-score is prioritized over the accuracy metric in the subsequent analyses [71]. The LRID and imbalance ratio remained constant for the 20, 30, and 40 second windows.

C. ANALYSIS OF FEATURES SELECTED OVER DIFFERENT WINDOW SIZES

Before evaluating the performance with the validation techniques, the SFS output was evaluated based on different epoch window sizes: 10, 20, 30, 40 seconds. Table 2 lists the average amount of features selected by SFS for varying windows sizes and tasks, which shows the SFS had optimal performance when 4-5 features were selected on average.

TABLE 1. Details of the multiclass datasets, M (±SD).

Task	Number of Observations (i.e., 10-sec Windows)	Class Observations	Imbalance Ratio	LRID
VR-ISS	76.3 (±41.6)	L: 25.8 (±17.8) M: 25.5 (±19.3) H: 25.0 (±16.4)	2.37 (±2.39)	11 (±17.4)
N-back	62.9 (±17.0)	L: 20.4 (±7.26) M: 20.5 (±6.0) H: 21.9 (±5.23)	1.52 (±1.71)	1.65 (±5.08)

TABLE 2. Number of features selected by SFS for each task.

Window	N-Back		VR-ISS	
	Mean	**STE**	**Mean**	**STE**
10 sec	4.6	0.2	4.6	0.1
20 sec	5.0	0.2	4.6	0.1
30 sec	4.8	0.2	4.7	0.1
40 sec	4.4	0.2	4.2	0.2

The frequency for VR-ISS and N-back features selected by SFS for varying epoch window sizes with an ABayes wrapper for 10-folds is illustrated in Fig. 7. Within the VR-ISS window sizes (10, 20, 30, 40 seconds), SBP mean (24%, 8%, 8%, 10%), SBP median (16%, 16%, 14%, 12%), DBP mean (25%, 11%, 10%, 5%), DBP median (8%, 18%, 19%, 12%) were selected for the most participants. The 30-seond window deviated from the other windows with different features being selected, such as DBP median becoming more prominent whereas the DBP and SBP means became less prominent for larger windows. The most frequent feature in any window was DBP mean (25%) followed by SBP mean (24%), which were in the 10-second window. Comparing the N-back window sizes, DBP mean (23%, 20%, 27%, 20%), SBP mean (28%, 19%, 9%, 6%), and followed by the DBP spectral density second coefficient (20%, 16%, 12%, 14%) and third coefficient (14%, 16%, 9%, 15%) were selected for the most participants. The most frequent feature in any window was SBP mean (28%) in the 10-second window and DBP mean (27%) in the 30-second window.

D. TASK AND WINDOW COMPARISON FOR ABAYES VALIDATION TECHNIQUES

The validation techniques were compared for the VR-ISS and N-back (Fig. 8). The window with the highest F1-score for the VR-ISS was 30-seconds (94%) for 10-Fold and 40 seconds (79%) for holdout. For the N-back task, the window with highest F1-score was 40-seconds (96%) for 10-Fold and 40-seconds (81%) for holdout.

E. CLASSIFIER COMPARISON FOR THE TASKS

The cross-validation and holdout results for the VR-ISS task with the various classifiers trained with physiological signal segments of different window sizes are summarized in Table 3. The highest F1-score for 10-Fold cross-validation was achieved with the ABayes classifier and highest F1-score for the holdout was with the SVM classifier. For 10-Fold, the best F1-score was for the ABayes model at 94% for a window size of 30-seconds. For the holdout, the F1-score was for the SVM model at 84% for a window size of 40 seconds. In comparison, ABayes was 5% lower than the SVM for 40-seconds holdout. The Decision Tree performed the worst out of the classifiers, with the lowest F1-score in every window for both 10-fold cross-validation and holdout.

The validation technique results for the N-back task are summarized in Table 4. The highest F1-score for the 10-Fold cross-validation was achieved with the Random Forest classifier and the highest for the holdout was achieved with the Decision Tree classifier. For 10-Fold, the highest F1-score was 98% with Random Forest for a window size of 30-seconds, which in contrast ABayes scored 10% lower. The second highest F1-score was a split between ABayes, Decision Tree, and Random Forest which all showed comparable performance of 96% for the 40-scecond window. SVM performed the worst for all cross-validation windows. For the holdout, the best F1-score was 84% for Decision Tree with a window size of 40-seconds, with ABayes being only 3% lower.

F. PERSONALIZED COMPARED TO A GENERALIZED APPROACH

The LOSO validation were compared between classifiers and window sizes for a subset of VR-ISS participants (Table 5). The classifier and window with the highest accuracy (62%) was the 30-second window using the Random Forest classifier.

V. DISCUSSION

The detection system was designed to select personalized time-series features that best describe the stress response for a given person, train the system with a post-hoc model and assess its effectiveness in classifying multiple levels of stress for real-time deployment. Like empirical studies, statistical techniques used for stress detection were judged in terms of objectivity, reliability, and validity [76].

Using a stress questionnaire, this experiment addressed the objectivity of its stress manipulation (i.e., ground truth reference of the stressor levels). The stress questionnaire showed that both the VR-ISS and N-back successfully separated participants' stress into three distinct levels. These same

FIGURE 7. **Frequency of SFS selection for each window size for VR-ISS and N-back during 10-fold cross-validation of the ABayes classifier. Features with less than 10% in every column were excluded from this figure for brevity. See appendix for the feature description and software package.**

stressor levels were previously reported to show significantly different physiological measures of stress (see [77]). This provides assurance that the machine learning classification is classified on distinct groups of data. The reliability of the system was assessed by comparing the classification methodology on two different tasks involving a standard laboratory cognitive task (N-back) compared to a complex job-specific task (VR-ISS), window size of the interval method, classifier validation techniques, and features selected by the wrapper. When comparing both tasks, the classifier performance was slightly better for the less-complex laboratory task of N-back. The features selected for each window varied, with

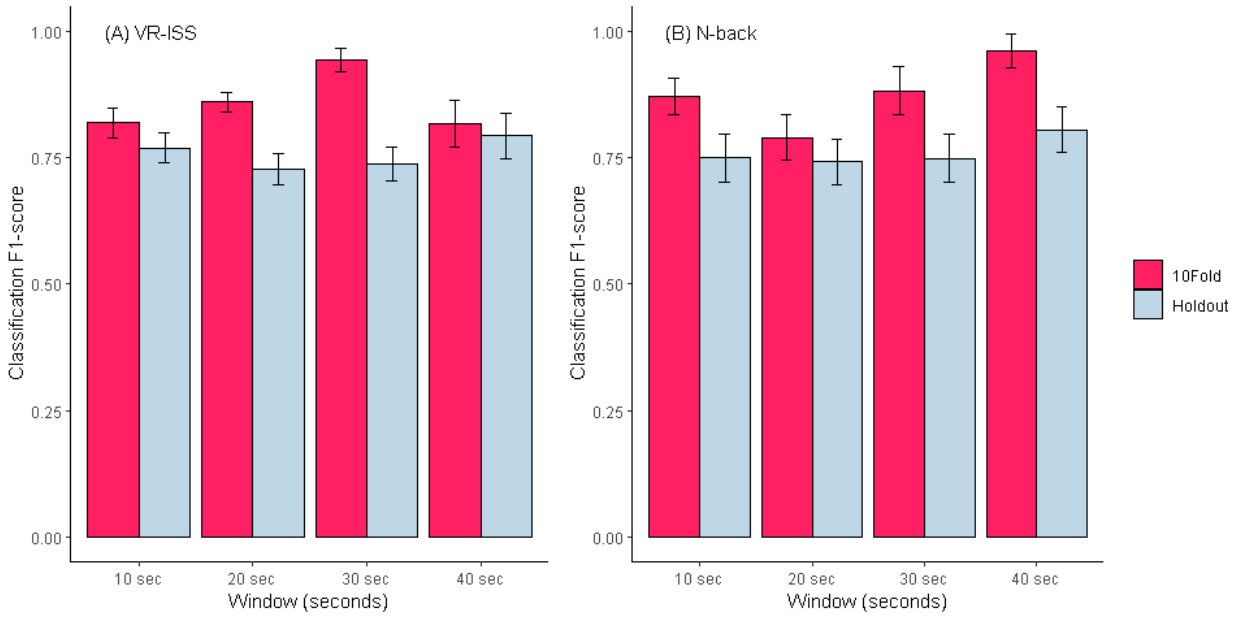


FIGURE 8. Validation technique comparison for ABayes during (A) VR-ISS, and (B) N-back tasks. Error bars in standard error.

TABLE 3. Results of the VR-ISS stress classification for different window sizes, classifiers, and validation techniques. Highest window F1-scores are highlighted.

		10Fold								Holdout							
		ABayes		DT		RF		SVM		ABayes		DT		RF		SVM	
		Mean	STE	Mean	STE	Mean	STE	Mean	STE	Mean	STE	Mean	STE	Mean	STE	Mean	STE
Window																	
10 sec	F1_score	0.82	0.03	0.80	0.03	0.84	0.03	0.81	0.03	0.77	0.03	0.73	0.03	0.77	0.03	0.71	0.04
	Accuracy	0.77	0.04	0.77	0.03	0.82	0.03	0.78	0.03	0.71	0.04	0.70	0.04	0.73	0.03	0.66	0.04
	Sensitivity	0.75	0.04	0.79	0.04	0.82	0.03	0.77	0.03	0.68	0.04	0.68	0.04	0.71	0.04	0.63	0.05
	Specificity	0.87	0.02	0.88	0.02	0.91	0.02	0.89	0.02	0.84	0.02	0.85	0.02	0.86	0.02	0.84	0.02
	Precision	0.81	0.04	0.80	0.04	0.85	0.03	0.80	0.03	0.66	0.05	0.72	0.04	0.74	0.04	0.71	0.05
20 sec	F1_score	0.86	0.02	0.90	0.03	0.87	0.03	0.91	0.02	0.73	0.03	0.73	0.03	0.74	0.04	0.70	0.04
	Accuracy	0.68	0.06	0.74	0.05	0.73	0.06	0.72	0.05	0.66	0.04	0.64	0.04	0.68	0.05	0.60	0.05
	Sensitivity	0.67	0.06	0.75	0.05	0.75	0.06	0.72	0.05	0.64	0.04	0.63	0.05	0.67	0.05	0.59	0.05
	Specificity	0.84	0.03	0.87	0.03	0.87	0.03	0.86	0.03	0.82	0.02	0.82	0.02	0.84	0.02	0.79	0.03
	Precision	0.72	0.06	0.76	0.05	0.76	0.06	0.74	0.06	0.65	0.04	0.68	0.05	0.73	0.04	0.63	0.05
30 sec	F1_score	0.94	0.03	0.89	0.03	0.90	0.03	0.93	0.03	0.74	0.03	0.77	0.05	0.72	0.03	0.78	0.03
	Accuracy	0.61	0.08	0.61	0.07	0.60	0.08	0.73	0.08	0.56	0.06	0.51	0.07	0.58	0.05	0.64	0.05
	Sensitivity	0.60	0.08	0.61	0.07	0.59	0.08	0.72	0.08	0.55	0.06	0.51	0.06	0.58	0.05	0.61	0.06
	Specificity	0.81	0.04	0.80	0.04	0.79	0.04	0.85	0.04	0.78	0.03	0.73	0.04	0.77	0.04	0.82	0.03
	Precision	0.62	0.08	0.62	0.07	0.60	0.08	0.74	0.08	0.57	0.06	0.52	0.07	0.61	0.06	0.67	0.06
40 sec	F1_score	0.82	0.06	0.88	0.05	0.93	0.03	0.88	0.04	0.79	0.05	0.67	0.05	0.79	0.04	0.84	0.04
	Accuracy	0.42	0.11	0.55	0.11	0.47	0.11	0.52	0.11	0.60	0.07	0.54	0.06	0.66	0.07	0.65	0.07
	Sensitivity	0.41	0.10	0.54	0.11	0.44	0.11	0.50	0.11	0.61	0.07	0.56	0.05	0.66	0.07	0.65	0.07
	Specificity	0.69	0.05	0.76	0.06	0.73	0.06	0.73	0.05	0.80	0.04	0.77	0.03	0.75	0.07	0.77	0.07
	Precision	0.43	0.11	0.54	0.11	0.47	0.11	0.54	0.11	0.61	0.07	0.58	0.06	0.69	0.07	0.70	0.08

the 10-20 sec windows having selected SBP mean and DBP mean more than the 30-40 second windows, suggesting that physiological timescale may influence feature classification performance. Since the training model was based on only one subject's data, using multiple individuals for each of those tasks increases the reliability that the findings can be repeated on future individuals. The results showed a possible data imbalance; therefore, the F1-score was used for classifier comparison. Results of cross-validation and holdout show optimal F1-scores ranging from 82-94% and 73-79%, respectively for the VR-ISS and 79-96% and 74-81%, respectively for the N-back. For validity, the ABayes algorithm was tested against other machine learning classifiers; the personalized model was compared to a generalized model's performance and compared to the results of other multi-class stress detection studies. Results show that the performance of traditional supervised machine learning classifiers are minor-to-moderately affected by indirect approximations through comparison to a benchmark optimal classifier (ABayes). Comparison of accuracy metrics between other studies had large caveats due to limited statistical reporting, despite the F1-score being a more accurate assessment of the classifier performance. The personalized model was found to perform better than a generalized model used on an optimized dataset. Overall, a personalized stress detection system had slightly lower accuracy at multi-class detection in comparison to other

TABLE 4. Results of the N-Back stress classification for different window sizes, classifiers, and validation techniques. Highest window F1-scores are highlighted.

		10Fold								Holdout							
		ABayes		DT		RF		SVM		ABayes		DT		RF		SVM	
		Mean	STE	Mean	STE	Mean	STE	Mean	STE	Mean	STE	Mean	STE	Mean	STE	Mean	STE
Window																	
10 sec	F1_score	0.87	0.04	0.90	0.03	0.86	0.03	0.86	0.04	0.75	0.05	0.82	0.04	0.75	0.05	0.76	0.04
	Accuracy	0.86	0.04	0.89	0.04	0.85	0.03	0.84	0.05	0.75	0.05	0.77	0.04	0.74	0.05	0.75	0.04
	Sensitivity	0.86	0.04	0.89	0.03	0.84	0.04	0.83	0.05	0.74	0.05	0.78	0.04	0.75	0.05	0.75	0.04
	Specificity	0.93	0.02	0.95	0.02	0.92	0.02	0.92	0.02	0.87	0.02	0.89	0.02	0.87	0.03	0.88	0.02
	Precision	0.92	0.02	0.92	0.03	0.88	0.03	0.85	0.05	0.77	0.05	0.79	0.04	0.77	0.05	0.77	0.05
20 sec	F1_score	0.79	0.05	0.88	0.04	0.93	0.03	0.86	0.04	0.74	0.05	0.76	0.06	0.78	0.04	0.74	0.05
	Accuracy	0.63	0.06	0.70	0.08	0.84	0.06	0.75	0.06	0.68	0.06	0.65	0.08	0.69	0.05	0.70	0.06
	Sensitivity	0.65	0.06	0.71	0.08	0.85	0.06	0.76	0.06	0.67	0.06	0.64	0.08	0.69	0.05	0.71	0.06
	Specificity	0.81	0.03	0.85	0.05	0.91	0.03	0.87	0.03	0.84	0.03	0.82	0.04	0.84	0.02	0.85	0.03
	Precision	0.65	0.06	0.73	0.08	0.86	0.06	0.77	0.07	0.68	0.06	0.65	0.09	0.71	0.05	0.74	0.06
30 sec	F1_score	0.88	0.05	0.95	0.03	0.98	0.02	0.96	0.04	0.75	0.05	0.73	0.05	0.75	0.05	0.81	0.05
	Accuracy	0.70	0.10	0.79	0.07	0.61	0.13	0.68	0.12	0.60	0.06	0.64	0.07	0.62	0.07	0.75	0.07
	Sensitivity	0.70	0.10	0.79	0.07	0.59	0.13	0.68	0.12	0.62	0.06	0.65	0.07	0.63	0.07	0.75	0.07
	Specificity	0.85	0.05	0.91	0.03	0.80	0.07	0.85	0.07	0.80	0.03	0.82	0.03	0.82	0.03	0.88	0.04
	Precision	0.70	0.10	0.80	0.07	0.61	0.13	0.68	0.12	0.62	0.07	0.68	0.07	0.70	0.07	0.78	0.08
40 sec	F1_score	0.96	0.04	0.96	0.04	0.96	0.04	0.85	0.06	0.81	0.05	0.84	0.06	0.83	0.05	0.80	0.06
	Accuracy	0.64	0.12	0.50	0.12	0.59	0.13	0.59	0.11	0.60	0.08	0.54	0.09	0.52	0.10	0.62	0.08
	Sensitivity	0.64	0.12	0.50	0.12	0.59	0.13	0.59	0.11	0.61	0.08	0.55	0.09	0.53	0.10	0.62	0.08
	Specificity	0.82	0.07	0.76	0.07	0.79	0.07	0.76	0.07	0.80	0.04	0.77	0.05	0.76	0.05	0.81	0.04
	Precision	0.64	0.12	0.50	0.12	0.59	0.13	0.59	0.11	0.62	0.09	0.56	0.10	0.57	0.10	0.64	0.09

TABLE 5. Accuracy of the VR-ISS generalized approach using LOSO validation.

Window	ABayes		DT		RF		SVM	
	Mean	**STE**	**Mean**	**STE**	**Mean**	**STE**	**Mean**	**STE**
10 sec	0.36	0.06	0.53	0.05	0.61	0.05	0.52	0.05
20 sec	0.46	0.06	0.53	0.04	0.57	0.06	0.55	0.06
30 sec	0.48	0.08	0.56	0.05	0.62	0.05	0.5	0.04
40 sec	0.41	0.07	0.56	0.03	0.56	0.05	0.51	0.05

studies, albeit this the VR-ISS accuracy is likely affected by the imbalanced data. The results from the F1-score performance suggest the personalized models can be beneficial and are suitable for deployment in a real-time stress detection system.

The advantage of comparing traditional classifiers to ABayes is the direct and transparent connection to probability modeling. This allows for conditional probability that may better represent the dataset as a whole since ABayes approximates the optimal Bayes solution by directly estimating the probability density with multivariate kernel density estimations rather than indirect approximations like softmax functions. This may benefit those seeking a benchmark classifier that wish to have strict density estimates while minimizing feature dependencies. Results show that the highest 10-fold cross validation performance for the VR-ISS across all windows and classifiers was 94% using an ABayes classifier with a window size of 30-seconds, suggesting the personalized approach performed well. The traditional classifiers underperformed or outperformed ABayes for specific window sizes, but generally were in a F1-score range of -11 to $+14$ of ABayes. This suggests that traditional classifiers were minor-to-moderately affected by indirect approximations.

When comparing ABayes with the traditional classifiers, a nuanced approach may be required to achieve the highest performing stress detection depending on the features, task, window size, and interpretability of the classifier's logic. ABayes, Random Forest, and SVM generally resulted in higher F1-score for the cross-validation and holdout with a VR-ISS task. For cross-validation with a VR-ISS task, the ABayes had the highest performance with an F1-score of 94% whereas the SVM had the highest holdout performance of 84%. In contrast, the N-back task results for Random Forest had the highest performance with an F1-score of 98% whereas the Decision Tree had the highest holdout performance of 84%. This suggests that both tasks had good stress detection performance; however, the data differed such that the best classifier varied between the tasks. The Decision Tree had the lowest performance compared to the other classifiers for the VR-ISS, but consistently was better for both validation techniques during the N-back.

Since individuals can respond differently between tasks/stressors, the detection was assessed between the N-back and VR-ISS. Results show the F1-score was slightly higher for the N-back than the VR-ISS. This is expected since the N-back is a more controlled laboratory task. The VR-ISS task is more complex and more closely matches the dynamic task demands faced in training of a real-world task. The highest N-back F1-score was for Random Forest with a 30-second window during 10-fold validation, resulting in 98% compared to 90% for VR-ISS with same parameters. The range of Random Forest between VR-ISS and N-Back is 82-94% and 79-96%, respectively. Similarly, the range of ABayes between VR-ISS and N-Back is 84-93% and 86-98%,

TABLE 6. Details of the multiclass datasets, M (±SD).

Reference	Levels	Classifier	Subjects	Sensors	Generalized/ Individualized	Accuracy (10Fold)
This study	3	Random Forest (30-sec)	7 (VR-ISS)	ECG, EDA, RSP, NIBP	Gen.	62%
This study	3	ABayes (10-sec)	27 (VR-ISS)	ECG, EDA, RSP, NIBP	Ind.	77%
This study	3	Random Forest (10-sec)	27 (VR-ISS)	ECG, EDA, RSP, NIBP	Ind.	82%
This study	3	ABayes (10-sec)	14 (N-Back)	ECG, EDA, RSP, NIBP	Ind.	86%
This study	3	Random Forest (30-sec)	14 (N-Back)	ECG, EDA, RSP, NIBP	Ind.	98%
[24]	4	SVM	20	ECG	Gen.	86.3%
[30]	3	Ada Boost	21	ECG, RSP	Ind.*	90.2%
[79]	3	SVM	4	ECG	Gen.	89.2%
[80]	3	Multi. Perceptron	17	ECG, EMG, RSP, EDA	Gen.	80.6%
[81]	3	Random Forest	21	ECG, EDA	Ind.	97.2%
[38]	3	SVM	34	ECG, EMG	Gen.	97.6%
[82]	4	SVM	30	BVP,EDA, ST	Gen.	86.3%
[83]	3	Decision Tree	17	ECG,EMG,EDA, RSP	Gen.	70.2%
[84]	3	Multi. Perceptron	28	EEG	Gen.	60.71%

* Generalized model, but calibrated to the individual using subjective stress scores.

respectively. Slightly higher accuracy was expected because the N-back is a more sustained and controlled stressor. Further, the N-back task is well validated at eliciting different levels of mental workload, specifically different levels of working memory, and physiological stress indices [48], [78]. In contrast, the VR-ISS was a more complex task involving a variety of stressors including, noise, task load, decreased visibility, and simulated physical threat. This trend also is present for holdout, which ABayes between VR-ISS and N-Back is 73-79% and 74-81%, respectively. The accuracy for both tasks was relatively close, suggesting this stress detection system may be robust in translation to other complex training tasks.

Directly comparing stress detection approaches is difficult due to varying factors in the pipeline development or differences in datasets. The generalized model was found to have lower accuracy (62%) than the personalized model (82%) on VR-ISS data (Table 6). While the approaches used different validation and features (because of the personalized used SFS), the comparison gives insight into how the models may perform on unseen data. Furthermore, the generalized had difficulty with high degree of individual variation in features. In contrast, the personalized approach was able to account for them by selecting features unique to that individual. Collecting homogenous data with more participants in the training dataset may improve the generalized model results.

In comparison to other research on multiclass stress detection, the 10-fold accuracy of ABayes for VR-ISS and N-back was 77% and 86%, while the Random Forest for VR-ISS and N-back was 82% and 98% (Table 6). The N-back accuracy values are some of the highest recorded values for multi-class stress detection. Meanwhile the ABayes values were lower. However, many of these studies used generalized classifiers, built models with predetermined features, and did not report an F1-score. As mentioned previously, the LRID metric suggested a high probability that the VR-ISS dataset was imbalanced which can affect the accuracy metric. Therefore, F1-score is favored for imbalanced datasets (reported as 94% for VR-ISS and 88% for N-back for ABayes with a 30-second window during 10-fold evaluation), yet most other researchers did not report a F1-score for comparison. Thus, Table 6 shows a biased accuracy VR-ISS and more comparable accuracy for N-back. Further, since physiological stress activation can vary between individuals, some studies found that individual stress detection models had higher classification accuracies than general models [81]. Individual models reported in Table 6 generally show higher 10-fold accuracy when comparing between studies. Lastly, the SFS wrapper used within this study's pipeline gives an added advantage as supported by a similar wrapper feature selection process that resulted in high accuracy [38]. Together, this suggests that model and evaluation parameters should be carefully considered when comparing machine learning metrics.

Analysis of the window size for feature selection offers insight into prominent time-series patterns. Both feature extraction packages used by the pipeline contain measures of mean, variance, linearity, stationarity, frequency, and entropy. The features selected were similar for the VR-ISS window sizes. The most prominent features were the mean, median, and power spectral density coefficients for the SBP and DBP signals. EDA mean and power spectral density coefficients were also prominent, but only for certain window sizes. These selected features fit the physiological narrative. SBP and EDA mean and median have been shown to be elevated during acute stressors [10], [85]. The power spectral density coefficients overlap the very-low and low frequency range (0.01-0.05 Hz). For SBP, this frequency range is associated with sympathetic activation of vascular tone [86]. Similarly,

the EDA low frequency range is associated with sympathetic activation from stress [87]. The selection of SBP and DBP features suggest that blood pressure may be an overlooked and underutilized stress biomarker and may be beneficial for stress detection. It is surprising that HRV frequency-domain features were not selected, considering that the HRV sympathetic measures are correlated to EDA [88]. Since these HRV calculations were verified against other HRV software, the most likely reason HRV features are not selected is due to significance testing criteria. The stress detection's feature selection used a one-way ANOVA to select features, however, a single-persons HRV metrics rarely rose to the level of significance between stressor levels. Longer durations for training data collection may increase the frequency that HRV metrics are selected.

The HRV finding reflects a common issue with most generalized detection systems, the omission of relevant individual data in favor of group averages [88]. There is a common understanding that the psychosocial stress response has a high degree of heterogeneity [89]. The average subjective stress of the manipulations showed distinct stressor levels (as do the HRV physiological response [77]), however, the variety of individual features selected by the personalized model indicate that there is meaningful variance in types of physiological systems activated that are not captured at the group level. Although HRV is a very common analysis in stress studies, it would not have been a reliable personalized feature in this case. Predetermining features for a model risks dataset bias (i.e., omission of relevant data used to train the model). These issues need to be addressed in the near term as the applied value of stress detection systems will be determined at the level of single participants.

As physiological systems act at different time-scales, the reliance on features changed as the analysis timeframe increased. The 30-40 second window had decreased selection of SBP, EDA, and RESP compared to the 10-20 second window. For the VR-ISS, the 30-40 second window showed reliance on DBP median, SBP mean absolute change, and SBP median. The N-back showed reliance in the 30-40 second windows on DBP median, DBP power spectral density coefficients and RESP median. This is somewhat expected as high frequency signals like EDA may be indicative of stress in shorter time intervals but become diluted over larger windows. These results show that the features selected are relatively different for the windows and tasks. Window sizes should be considered with selecting features for stress detection but may show varied responses for different tasks that may induce different physical and psychological demands on the individual. This suggest the generalized stress detection systems may not be robust for changing stressful scenarios due to each task evoking a different physiological stress response.

Past stress detection approaches largely use predetermined features or create a generalized system without consideration for the time series nature of those features, but also that those physiological signals may have different responses (i.e., signal strength through neural activation) in different individuals. The results show that many different features may be selected across individuals. Further, the low selection rates of features such as heart rate, EDA, and RMSSD indicate some individuals may have strong enough responses to discriminate multiple levels of stress using those features, while other individuals may lack the magnitude of change required for detection, leading to lower classification accuracy. A system with those predetermined features may be limiting the generalizability of the detection system to a larger population. Future stress detection systems will need to address individual differences in stress responses in order to enhance detection capabilities.

The wrapper method was necessary to select features that could be used to personalize the model with features best discriminated stress for given window sizes, and subsequently deployed to test real-time data in an adaptive system [14], [15], [16]. Without a wrapper, features would have to be predetermined and would generalize to a broad population when the system is deployed. However, they would neglect important individual differences in the physiological stress response. The simplest way to implement an optimized pipeline is to have the classifier that is within the wrapper match the classifier that is used during the validation techniques. Other potential solutions are to use wrapper-based decision trees to combine multiple classifiers to select mutually agreed relevant features [90]. It may be beneficial for stress detection systems to continue to develop new methods of personalization that account for differences in the physiological stress responses between individuals.

While this study primarily used physiological measures to predict stress, there may also be some benefit to including other peripheral indicators of physiology or cognitive state, such as demographics, behavior, gestures, eye movements, and speech. Past research has shown that when physiological activity is measured over a long term, factors including demographic and lifestyle (e.g., age, sex, physical activity, alcohol use) show association to HRV [91]. Similarly, behavior and gestures may indicate an emotional response from stress, which have been previous use in stress detection [92]. While pupil dilation is a well-known response controlled by the autonomic nervous system, eye movements (e.g., glare, saccade) can be indicators of attention tunneling from heightened threat appraisal [93]. Lastly, the linguistic and para-linguistic features of speech and be used to detect stress [31]. As detection methods continue to improve, stress detection systems will increasingly move toward mobile and less intrusive sensors.

The feasibility of the approach described in the study is largely dependent on the sensors, signal quality, and type of stress being detected. When designing systems, there must be a balance between mobility and the system sensors ability to capture enough variability to model the physiological stress response. The stress response is complex and involves

multiple systems (i.e., neural, neuroendocrine, endocrine), some that respond to acute stress on sort timescales and some that measure chronic stress over longer time frames. Further, as noted by a 2022 meta-analysis [19], physiological systems are differentially activated by types of stressors (e.g., acute social stressors are more likely to increase the hormone cortisol). Since model training data obtains the best predictive results when it captures similar patterns as the testing data [94], the variation in both the psychological and physiological stress response due to task context suggests that models trained using stress signals from one task (e.g., N-back) made not represent the stress in a different task (e.g., VR-ISS). For this study which developed acute stress detection for training in virtual reality, the individual models were trained and tested using the same task (i.e., N-back trained models tested N-back data). The low amount of physical movement allowed for multiple sensors (e.g., ECG, GSR, BP) while also providing a high degree of certainty in the stress response measurement. Further, the approach using real-time preprocessing on signals and feature selection wrapper was able to exclude or rectify features that might be affected by signal artifacts or individual differences in the physiological response (e.g., less neural activation than normal). While this study benefited from high-quality sensors that were able to send the data to the computer via Bluetooth, future research is needed to use the personalized approach with embedded and/or lower quality sensors through enhancement of the preprocessing and feature selection criteria.

The experimental results reflect the system performance in real time because the real-time and post-hoc systems use the same preprocessing, feature selection, and feature extraction under the same experimental interventions. However, the classification performance will deteriorate over time in a real-time system. The greatest limitation of the real-time system is the inability to retrain the machine learning model or calibrate over time due to changes in the appraisal of a stressor (i.e., reappraisal) or the physiological habituation of biomarkers over repeated exposure. In practice, this makes model validation of real-time data for stress detection infeasible and erroneous unless the system can compensate for individual changes over time. For example, systems that measure acute or chronic stress through repeated sampling of salivary cortisol [95] or sweat cortisol levels [96], could become continuous stress detection systems by adjusting the models based on the body's adaptive physiological changes in cortisol response and reactivity [97] which may become dysregulated (blunted or sensitized) from exposure to stress. Since the stress response is largely dependent on psychological appraisal, model calibration has been attempted using subjective rating [30]. However, additional challenges may arise from attempting to use subjective questionnaires to calibrate the models, as subjective stress metrics have been shown to have a poor correlation to the physiological stress response [98]. Both the reported real-time system and future stress detection systems need a way to occasionally adjust their models to account for how the individual changes over time.

While this study primarily used young healthy subjects, physiological systems undergo age-related changes that may influence the feature selection. For example, there is an age-related increase in SBP reactivity and parasympathetic withdrawal to acute stress [99]. Further work is needed to determine the impact of these changes on stress detection.

Another possible limitation is wrapper overfitting of the models due to highly correlated variables. The SFS wrapper fit the model by selecting a combination of features that resulted in the highest accuracy. However, the wrapper did not account for the correlation between features. If all the features are selected from the same sensor, it not only neglects important physiological responses in other bodily systems but also places undue reliance on the sole sensor working correctly [100]. Another limitation is that some estimation error could have been caused by underspecification, which occurs when the training process has multiple predictors (e.g., feature structures) that appear equal, but have divergent performance when deployed [101]. In this study, the SFS wrapper selected the feature subset with the highest wrapper accuracy but chose the subset with the least amount of features if multiple subsets had equal maximal performance. These subsets may have had different performance during cross-validation or holdout, and further research is needed to evaluate performance when deployed. Another potential limitation is the ABayes will only be optimal in the case that the features follow Gaussian distributions because the classifier used a Univariate Gaussian kernel smoothing parameter [102]. Further, the distributions obtained by some features, like phasic and tonic EDA, can on occasion be non-Gaussian. Thus, enhancing the ability of ABayes to handle non-parametric data may further its utility as a benchmark. Finally, it is unclear if different classifiers implemented within the features selection wrapper may have varying feature choices depending on window size (i.e., the prominence of selecting features for increasing window sizes). Future work should evaluate if there is a significant relationship between classifier wrapper and epoch length.

VI. CONCLUSION

To address the challenges of vast differences between individual stress response, the time-series nature of physiological signals, this research evaluated the objectivity, reliability, and validity of a real-time stress detection system using a personalized time-series interval approach. The simple and complex tasks were able to achieve distinct levels of stress enabling their use as machine learning ground truth. Analysis of the window sizes provided insight into which sensors/features were useful for varying time-intervals. The personalized model was found to have better performance than a generalized model. Furthermore, it evaluated the effect of indirect approximations by supervised machine learning classifiers evaluated against a benchmark optimal classifier, ABayes. It was found that indirect approximations

TABLE 7. List of all features included in the feature extraction, grouped by signal and listed in alphabetical order.

#	Feature	Abbreviation	Reference/Package
Extracted from all Biopac signals (HR,RMSSD,pNN50,RSP,NIBP,SBP,DBP,EDA,EDAtonic,EDAphasic)			
1	Abs Energy	abs_energy	Tsfresh
2	Absolute sum of changes	abs_sum_changes	Tsfresh
3	Augmented Dickey Fuller	adfpValue	Tsfresh
4	Autocorrelation	Agg_AutoC	Tsfresh
5	Autocorrelation, first 1/e crossing	CO_f1ecac	Catch22
6	Autocorrelation, fist minimum	CO_FirstMin_ac	Catch22
7	Automutual info, m=2, τ=5	CO_HistogramAMI_even_2_5	Catch22
8	Automutual, first minimum	IN_AutoMutualInfoStats_40_gaussian_fmmi	Catch22
9	Binned entropy	binned_entropy	Tsfresh
10	C3 non-linearity measure	c3	Tsfresh
11	Change in correlation length after iterative differencing	FC_LocalSimple_mean1_tauresrat	Catch22
12	Complexity-invariant distance	cid_ce	Tsfresh
13	Count above mean	GTmean	Tsfresh
14	Count below mean	LTmean	Tsfresh
15	Cross Correlation	CC	[44]
16	Exponential fit to successive distances in 2-d embedding space	CO_Embed2_Dist_tau_d_expfit_meandiff	Catch22
17	FFT aggregated spectral variance	fft_agg_var	Tsfresh
18	FFT real coefficients	rfft_real	Tsfresh
19	First location of maximum	first_loc_max	Tsfresh
20	First location of minimum	first_loc_min	Tsfresh
21	Kurtosis	Kurt	Tsfresh
22	Last location of maximum	last_loc_max	Tsfresh
23	Last location of minimum	last_loc_min	Tsfresh
24	Linear trend slope	linear_slope	Tsfresh
25	Longest strike above mean	longest_strike_above	Tsfresh
26	Longest strike below mean	longest_strike_below	Tsfresh
27	Mean	mean	Tsfresh
28	Mean absolute change	mean_abs_change	Tsfresh
29	Mean change	mean_change	Tsfresh
30	Mean second derivative central	mean_2nd	Tsfresh
31	Median	median	Tsfresh
32	Mode of distribution (5,10-bin histo)	DN_HistogramMode_5,10	Catch22
33	Number of crossings mean	num_cross_mean	Tsfresh
34	Partial Autocorrelations (1 lag)	Agg_PAutoC	Tsfresh
35	Percent of reoccurring data points to all data points	per_reocurr_dtp	Tsfresh
36	Percent of reoccurring values to all values	per_reocurr_val	Tsfresh
37	Ratio value number to time series length	ratio_val_totime_series	Tsfresh
38	Rolling 3-sample mean forecasting error	FC_LocalSimple_mean3_stderr	Catch22
39	Skewness	skew	Tsfresh
40	Periodicity measure	PD_PeriodicityWang_th0_01	Catch22
41	Power spectrum – Fourier, centroid	SP_Summaries_welch_rect_centroid	Catch22
42	Power spectrum – Fourier, total power of lowest fifth frequency	SP_Summaries_welch_rect_area_5_1	Catch22

TABLE 7. *(Continued.)* **List of all features included in the feature extraction, grouped by signal and listed in alphabetical order.**

43	Proportion of slower timescale fluctuations that scale with DFA (50% sampling)	SC_FluctAnal_2_dfa_50_1_2_logi_prop_r1	Catch22
44	Proportion of slower timescale fluctuations that scale with linearly rescaled range fits	SC_FluctAnal_2_rsrangefit_50_1_logi_prop_r1	Catch22
4546	Shannon entropy of two successive letters in equiprobable 3-letter symbolization	SB_MotifThree_quantile_hh	Catch22
47	Standard deviation	stddev	Tsfresh
48	Standard deviation of successive differences	SDSD	[44]
49	Successive differences exceeding 0.04σ	MD_hrv_classic_pnn40	Catch22
50	Successive differences longest period of decreases	SB_BinaryStats_diff_longstretch0	Catch22
51	Sum of reoccurring data points	sum_reoccuring_dpt	Tsfresh
52	Sum of reoccurring values	sum_reoccuring_val	Tsfresh
53	Sum of squares	SS	Tsfresh
54	Sum values	sum_val	Tsfresh
55	Time intervals between events above mean	DN_OutlierInclude_p_001_mdrmd	Catch22
56	Time intervals between events below mean	DN_OutlierInclude_n_001_mdrmd	Catch22
57	Time reversal asymmetry statistic	time_reversal	Tsfresh
58	Time-reversibility statistic	CO_trev_1_num	Catch22
59	Trace of covariance of transition matrix between symbols in 3-letter alphabet	SB_TransitionMatrix_3ac_sumdiagcov	Catch22
60	Variance	variance	Tsfresh
Extracted from all ECG Biopac signal			
61	Baseline width of the RR interval histogram	TINN	HRVTool
62	Heart rate	HR	HRVTool
63	High Frequency power	HF	HRVTool
64	Low Frequency power	LF	HRVTool
65	Peak Freq. of Low Freq. Band	pLF	HRVTool
66	Peak Freq. of High Freq. Band	pHF	HRVTool
67	Percent of R peaks in ECG that differ more than 50 millisecond	pNN50	HRVTool
68	Percent of R peaks in ECG that differ more than 20 millisecond	pNN20	HRVTool
69	Poincaré plot standard deviation perpendicular the line of identity	SD1	HRVTool
70	Poincaré plot standard deviation along the line of identity	SD2	HRVTool
71	Ratio of SD1-to-SD2	SD1SD2ratio	HRVTool
72	Ratio of Low-High Frequency power	LFHFratio	HRVTool
73	Root Mean Square of Successive Difference of RR interval	RMSSD	HRVTool
74	Standard deviation of successive differences	SDSD	HRVTool
75	Standard deviation of NN intervals	SDNN	HRVTool
76	Triangular index from the interval histogram	TRI	HRVTool
77	Very Low Frequency power	VLF	HRVTool

can have a minor-to-moderate effect on classifier performance (-11% to +14% of ABayes). The current findings suggest that a personalized system provides promising performance when compared to past research on multi-class stress detection. Researchers should be careful about the selection of HMIs, sensors, and features for models, as they may not account for inter and intra- individual differences in stress physiology. Future work will further investigate these personalized stress detection systems with the aim of implementing approaches that account for temporal changes in the individual stress response and physiological signals.

TABLE 8. List of all features included in the generalized model and listed in alphabetical order.

#	Feature
1	DBP_abs_energy
2	DBP_abs_sum_changes
3	DBP_c3
4	DBP_mean
5	EDAtonic_abs_sum_changes
6	EDAtonic_c3
7	EDAtonic_cid_ce
8	EDAtonic_longest_strike_below
9	EDAtonic_mean
10	EDAtonic_skew
11	EDAtonic_stddev
12	EDAtonic_sum_reoccuring_dpt
13	EDAtonic_sum_reoccuring_val
14	HeartRate_c3
15	HeartRate_cid_ce
16	HeartRate_longest_strike_below
17	HeartRate_mean
18	HeartRate_skew
19	HeartRate_stddev
20	HeartRate_sum_reoccuring_val
21	RMSSD_abs_sum_changes
22	RMSSD_c3
23	RMSSD_cid_ce
24	RMSSD_longest_strike_above
25	RMSSD_longest_strike_below
26	RMSSD_mean
27	RMSSD_skew
28	RMSSD_stddev
29	RMSSD_sum_reoccuring_dpt
30	RMSSD_sum_reoccuring_val
31	SBP_abs_energy
32	SBP_c3
33	SBP_cid_ce
34	SBP_longest_strike_above
35	SBP_longest_strike_below
36	SBP_mean
37	SBP_skew
38	SBP_stddev
39	SBP_sum_reoccuring_val

APPENDIX

The features extracted from the physiological signals are listed in Table 7. All features included in the generalized model are listed in Table 8.

ACKNOWLEDGMENT

The authors thank Robin Gillund, Silvia Verhofste, Matthew Kreul, and Kelly Thompson for their laboratory assistance with research participants. The authors thank Pete Evans, Grant Leacox, Peter Carlson, and Robert Slezak for their help developing the VR-ISS and fire equipment models.

REFERENCES

[1] J. E. Driskell, E. Salas, J. H. Johnston, and T. N. Wollert, *Stress Exposure Training: An Event-Based Approach* (Performance Under Stress). London, U.K.: Ashgate, 2008, pp. 271–286.
[2] I. Barshi and D. L. Dempsey, "Risk of performance errors due to training deficiencies: Evidence report," Nat. Aeronaut. Space Admin. (NASA), NASA Johnson Space Center, Houston, TX, USA, Tech. Rep., JSC-CN-35755, 2016.
[3] M. Gjoreski, M. Luštrek, M. Gams, and H. Gjoreski, "Monitoring stress with a wrist device using context," *J. Biomed. Inform.*, vol. 73, pp. 159–170, Sep. 2017, doi: 10.1016/j.jbi.2017.08.006.
[4] M. Zahabi and A. M. A. Razak, "Adaptive virtual reality-based training: A systematic literature review and framework," *Virtual Reality*, vol. 24, no. 4, pp. 725–752, Dec. 2020, doi: 10.1007/s10055-020-00434-w.
[5] Y. S. Can, B. Arnrich, and C. Ersoy, "Stress detection in daily life scenarios using smart phones and wearable sensors: A survey," *J. Biomed. Informat.*, vol. 92, Apr. 2019, Art. no. 103139, doi: 10.1016/j.jbi.2019.103139.
[6] A. O. Akmandor and N. K. Jha, "Keep the stress away with SoDA: Stress detection and alleviation system," *IEEE Trans. Multi-Scale Comput. Syst.*, vol. 3, no. 4, pp. 269–282, Oct. 2017, doi: 10.1109/tmscs.2017.2703613.
[7] M. Verleysen and D. Franaois, "The curse of dimensionality in data mining and time series prediction," in *Proc. Int. Work-Conf. Artif. Neural Netw.* Berlin, Germany: Springer, 2005, pp. 758–770, doi: 10.1007/11494669_93.
[8] S. Tong and D. Koller, "Bayes optimal hyperplanes? Maximal margin hyperplanes," in *Proc. IJCAI*, 1999, pp. 1–5.
[9] I. Rish, "An empirical study of the naive Bayes classifier," in *Proc. IJCAI Workshop Empirical Methods Artif. Intell.*, 2001, vol. 3, no. 22, pp. 41–46.
[10] F. Shaffer, R. McCraty, and C. L. Zerr, "A healthy heart is not a metronome: An integrative review of the heart's anatomy and heart rate variability," *Frontiers Psychol.*, vol. 5, p. 1040, Sep. 2014, doi: 10.3389/fpsyg.2014.01040.
[11] B. Kim, Y.-S. Jeong, and M. K. Jeong, "New multivariate kernel density estimator for uncertain data classification," *Ann. Oper. Res.*, vol. 303, nos. 1–2, pp. 413–431, Aug. 2021.
[12] E. Smets, W. De Raedt, and C. Van Hoof, "Into the wild: The challenges of physiological stress detection in laboratory and ambulatory settings," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 463–473, Mar. 2019, doi: 10.1109/JBHI.2018.2883751.
[13] D. Jones and S. Dechmerowski, "Measuring stress in an augmented training environment: Approaches and applications," in *Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience*. Berlin, Germany: Springer, 9744, pp. 23–33, 2016, doi: 10.1007/978-3-319-39952-2_3.
[14] T. T. Finseth, "Adaptive virtual reality stress training for spaceflight emergency procedures," Ph.D. dissertation, Aerosp. Eng., Iowa State Univ., 2021.
[15] T. Finseth, M. C. Dorneich, N. Keren, W. D. Franke, S. Vardeman, J. Segal, A. Deick, E. Cavanah, and K. Thompson, "The effectiveness of adaptive training for stress inoculation in a simulated astronaut task," in *Proc. Human Factors Ergonom. Soc. Annu. Meeting*, Baltimore, MD, USA, 2021, pp. 1541–1545.
[16] T. Finseth, M. C. Dorneich, N. Keren, W. D. Franke, and S. Vardeman, "Training for stressful operations using adaptive systems: Conceptual approaches and applications," in *Proc. Interservice/Industry Training, Simul. Educ. Conf. (I/ITSEC)*, Orlando, FL, USA, 2021, pp. 1–13.
[17] H. F. Posada-Quintero and K. H. Chon, "Innovations in electrodermal activity data collection and signal processing: A systematic review," *Sensors*, vol. 20, no. 2, p. 479, Jan. 2020, doi: 10.3390/s20020479.
[18] B. M. Kudielka, D. H. Hellhammer, and S. Wust, "Why do we respond so differently? Reviewing determinants of human salivary cortisol responses to challenge," *Psychoneuroendocrinology*, vol. 34, no. 1, pp. 2–18, Jan. 2009, doi: 10.1016/j.psyneuen.2008.10.004.

[19] L. V. Dammen, T. T. Finseth, B. H. McCurdy, N. P. Barnett, R. A. Conrady, A. G. Leach, A. F. Deick, A. L. Van Steenis, R. Gardner, B. L. Smith, A. Kay, and E. A. Shirtcliff, "Evoking stress reactivity in virtual reality: A systematic review and meta-analysis," *Neurosci. Biobehavioral Rev.*, vol. 138, Jul. 2022, Art. no. 104709, doi: 10.1016/j.neubiorev.2022.104709.
[20] S. L. Bowers, S. D. Bilbo, F. S. Dhabhar, and R. J. Nelson, "Stressor-specific alterations in corticosterone and immune responses in mice," *Brain, Behav., Immunity*, vol. 22, no. 1, pp. 105–113, Jan. 2008, doi: 10.1016/j.bbi.2007.07.012.
[21] U. Reimer, E. Laurenzi, E. Maier, and T. Ulmer, "Mobile stress recognition and relaxation support with SmartCoping: User-adaptive interpretation of physiological stress parameters," in *Proc. 50th Hawaii Int. Conf. Syst. Sci.*, 2017, pp. 3497–3606, doi: 10.24251/HICSS.2017.435.
[22] R. R. Singh, S. Conjeti, and R. Banerjee, "Assessment of driver stress from physiological signals collected under real-time semi-urban driving scenarios," *Int. J. Comput. Intell. Syst.*, vol. 7, no. 5, p. 909, 2014, doi: 10.1080/18756891.2013.864478.
[23] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 2, pp. 156–166, Jun. 2005, doi: 10.1109/TITS.2005.848368.
[24] G. Tartarisco, N. Carbonaro, A. Tonacci, G. M. Bernava, A. Arnao, G. Crifaci, P. Cipresso, G. Riva, A. Gaggioli, D. De Rossi, A. Tognetti, and G. Pioggia, "Neuro-fuzzy physiological computing to assess stress levels in virtual reality therapy," *Interacting Comput.*, vol. 27, no. 5, pp. 521–533, Sep. 2015, doi: 10.1093/iwc/iwv010.
[25] S. Betti, R. M. Lova, E. Rovini, G. Acerbi, L. Santarelli, M. Cabiati, S. D. Ry, and F. Cavallo, "Evaluation of an integrated system of wearable physiological sensors for stress monitoring in working environments by using biological markers," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 8, pp. 1748–1758, Aug. 2018, doi: 10.1109/tbme.2017.2764507.
[26] F.-T. Sun, C. Kuo, H.-T. Cheng, S. Buthpitiya, P. Collins, and M. Griss, "Activity-aware mental stress detection using physiological sensors," in *Proc. Int. Conf. Mobile Comput., Appl., Services*, 2012, pp. 211–230, doi: 10.1007/978-3-642-29336-8_12.
[27] K. Hovsepian, M. Al'Absi, E. Ertin, T. Kamarck, M. Nakajima, and S. Kumar, "cStress: Towards a gold standard for continuous stress assessment in the mobile environment," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2015, pp. 493–504, doi: 10.1145/2750858.2807526.
[28] R. Martinez, E. Irigoyen, A. Arruti, J. I. Martin, and J. Muguerza, "A real-time stress classification system based on arousal analysis of the nervous system by an F-state machine," *Comput. Methods Programs Biomed.*, vol. 148, pp. 81–90, Sep. 2017, doi: 10.1016/j.cmpb.2017.06.010.
[29] V. Alexandratos, M. Bulut, and R. Jasinschi, "Mobile real-time arousal detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4394–4398, doi: 10.1109/icassp.2014.6854432.
[30] K. Plarre, A. Raij, S. M. Hossain, A. A. Ali, M. Nakajima, M. Al'Absi, E. Ertin, T. Kamarck, S. Kumar, M. Scott, and D. Siewiorek, "Continuous inference of psychological stress from sensory measurements collected in the natural environment," in *Proc. 10th ACM/IEEE Int. Conf. Inf. Process. Sensor Netw.*, Apr. 2011, pp. 97–108.
[31] K. M. Feigh, M. C. Dorneich, and C. C. Hayes, "Toward a characterization of adaptive systems: A framework for researchers and system designers," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 54, no. 6, pp. 1008–1024, Dec. 2012, doi: 10.1177/0018720812443983.
[32] X. Gu, Z. Cao, A. Jolfaei, P. Xu, D. Wu, T.-P. Jung, and C.-T. Lin, "EEG-based brain-computer interfaces (BCIs): A survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 5, pp. 1645–1666, Sep. 2021, doi: 10.1109/TCBB.2021.3052811.
[33] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, "Review on psychological stress detection using biosignals," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 440–460, Jan. 2022, doi: 10.1109/taffc.2019.2927337.
[34] V. Novak, P. Novak, J. de Champlain, A. R. Le Blanc, R. Martin, and R. Nadeau, "Influence of respiration on heart rate and blood pressure fluctuations," *J. Appl. Physiol.*, vol. 74, no. 2, pp. 617–626, Feb. 1993, doi: 10.1152/jappl.1993.74.2.617.
[35] B. D. Fulcher, "Feature-based time-series analysis," in *Feature Engineering for Machine Learning and Data Analytics*. Boca Raton, FL, USA: CRC Press, 2018, pp. 87–116.
[36] A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh, "The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances," *Data Mining Knowl. Discovery*, vol. 31, no. 3, pp. 606–660, May 2017, doi: 10.1007/s10618-016-0483-9.
[37] A. Zarei and B. M. Asl, "Automatic classification of apnea and normal subjects using new features extracted from HRV and ECG-derived respiration signals," *Biomed. Signal Process. Control*, vol. 59, May 2020, Art. no. 101927, doi: 10.1016/j.bspc.2020.101927.
[38] S. Pourmohammadi and A. Maleki, "Stress detection using ECG and EMG signals: A comprehensive study," *Comput. Methods Programs Biomed.*, vol. 193, Sep. 2020, Art. no. 105482, doi: 10.1016/j.cmpb.2020.105482.
[39] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers Public Health*, vol. 5, p. 258, Sep. 2017, doi: 10.3389/fpubh.2017.00258.
[40] A. Saeed, S. Trajanovski, M. Van Keulen, and J. Van Erp, "Deep physiological arousal detection in a driving simulator using wearable sensors," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2017, pp. 486–493, doi: 10.1109/icdmw.2017.69.
[41] B. Hidasi and C. Gaspar-Papanek, "ShiftTree: An interpretable model-based approach for time series classification," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Berlin, Germany: Springer, 2011, pp. 48–64, doi: 10.1007/978-3-642-23783-6_4.
[42] J. Braithwaite, D. Watson, and R. Jones, "A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments," *Psychophysiology*, vol. 49, no. 1, pp. 1017–1034, 2013.
[43] H. M. Burke, M. C. Davis, C. Otte, and D. C. Mohr, "Depression and cortisol responses to psychological stress: A meta-analysis," *Psychoneuroendocrinology*, vol. 30, no. 9, pp. 846–856, Oct. 2005, doi: 10.1016/j.psyneuen.2005.02.010.
[44] E. Campbell, A. Phinyomark, and E. Scheme, "Feature extraction and selection for pain recognition using peripheral physiological signals," *Frontiers Neurosci.*, vol. 13, p. 437, May 2019, doi: 10.3389/fnins.2019.00437.
[45] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2002, pp. 694–699, doi: 10.1145/775047.775151.
[46] T. T. Finseth, N. Keren, M. C. Dorneich, W. D. Franke, C. C. Anderson, and M. C. Shelley, "Evaluating the effectiveness of graduated stress exposure in virtual spaceflight hazard training," *J. Cognit. Eng. Decis. Making*, vol. 12, no. 4, pp. 248–268, Dec. 2018, doi: 10.1177/1555343418775561.
[47] T. Finseth, M. C. Dorneich, N. Keren, W. D. Franke, and S. Vardeman, "Designing training scenarios for stressful spaceflight emergency procedures," in *Proc. AIAA/IEEE 39th Digit. Avionics Syst. Conf. (DASC)*, Oct. 2020, pp. 1–10, doi: 10.1109/dasc50938.2020.9256403.
[48] C. Herff, D. Heger, O. Fortmann, J. Hennrich, F. Putze, and T. Schultz, "Mental workload during N-back task—Quantified in the prefrontal cortex using fNIRS," *Frontiers Hum. Neurosci.*, vol. 7, p. 935, Jan. 2014, doi: 10.3389/fnhum.2013.00935.
[49] B. S. Hantono, L. E. Nugroho, and P. I. Santosa, "Mental stress detection via heart rate variability using machine learning," *Int. J. Electr. Eng. Informat.*, vol. 12, no. 3, pp. 431–444, Sep. 2020, doi: 10.15676/ijeei.2020.12.3.3.
[50] A. Hernández-Sabaté, J. Yauri, P. Folch, M. À. Piera, and D. Gil, "Recognition of the mental workloads of pilots in the cockpit using EEG signals," *Appl. Sci.*, vol. 12, no. 5, p. 2298, Feb. 2022, doi: 10.3390/app12052298.
[51] R. Abbuhl, S. Gass, and A. Mackey, "Experimental research design," in *Research Methods in Linguistics*, R. J. Podesva and D. Sharma, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2014, pp. 116–134.
[52] *International Space Station, Emergency Procedures 1A: Depress, Fire, Equipment Retrieval (No. JSC-48566)*, National Aeronautics and Space Administration (NASA), Houston, TX, USA, 2013.
[53] G. S. Shields, M. A. Sazma, and A. P. Yonelinas, "The effects of acute stress on core executive functions: A meta-analysis and comparison with cortisol," *Neurosci. Biobehavioral Rev.*, vol. 68, pp. 651–668, Sep. 2016, doi: 10.1016/j.neubiorev.2016.06.038.
[54] R. R. Singh, S. Conjeti, and R. Banerjee, "A comparative evaluation of neural network classifiers for stress level analysis of automotive drivers using physiological signals," *Biomed. Signal Process. Control*, vol. 8, no. 6, pp. 740–754, Nov. 2013, doi: 10.1016/j.bspc.2013.06.014.
[55] N. Sharma and T. Gedeon, "Objective measures, sensors and computational techniques for stress recognition and classification: A survey," *Comput. Methods Programs Biomed.*, vol. 108, no. 3, pp. 1287–1301, Dec. 2012, doi: 10.1016/j.cmpb.2012.07.003.
[56] R. Carpenter, "A review of instruments on cognitive appraisal of stress," *Arch. Psychiatric Nursing*, vol. 30, no. 2, pp. 271–279, Apr. 2016, doi: 10.1016/j.apnu.2015.07.002.
[57] HTC Corporation. (2016). *Vive-Home*. [Online]. Available: https://www.htcvive.com/ca/

[58] S. Greene, H. Thapliyal, and A. Caban-Holt, "A survey of affective computing for stress detection: Evaluating technologies in stress detection for better health," *IEEE Consum. Electron. Mag.*, vol. 5, no. 4, pp. 44–56, Oct. 2016, doi: 10.1109/mce.2016.2590178.

[59] P. Karthikeyan, M. Murugappan, and S. Yaacob, "Detection of human stress using short-term ECG and HRV signals," *J. Mech. Med. Biol.*, vol. 13, no. 2, Apr. 2013, Art. no. 1350038, doi: 10.1142/s0219519413500383.

[60] S. Z. Bong, M. Murugappan, and S. Yaacob, "Methods and approaches on inferring human emotional stress changes through physiological signals: A review," *Int. J. Med. Eng. Informat.*, vol. 5, no. 2, p. 152, 2013, doi: 10.1504/ijmei.2013.053332.

[61] B. S. Zheng, M. Murugappan, and S. Yaacob, "Human emotional stress assessment through heart rate detection in a customized protocol experiment," in *Proc. IEEE Symp. Ind. Electron. Appl.*, Sep. 2012, pp. 293–298, doi: 10.1109/isiea.2012.6496647.

[62] B. Figner and R. O. Murphy, "Using skin conductance in judgment and decision making research," in *A Handbook of Process Tracing Methods for Decision Research*, M. Schulte-Mecklenbeck, A. Kuehberger, and R. Ranyard, Eds. New York, NY, USA: Psychology Press, 2011, pp. 163–184.

[63] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr, "Time series FeatuRe extraction on basis of scalable hypothesis tests (tsfresh—A Python package)," *Neurocomputing*, vol. 307, pp. 72–77, Sep. 2018, doi: 10.1016/j.neucom.2018.03.067.

[64] C. H. Lubba, S. S. Sethi, P. Knaute, S. R. Schultz, B. D. Fulcher, and N. S. Jones, "*catch22*: CAnonical time-series CHaracteristics," *Data Mining Knowl. Discovery*, vol. 33, no. 6, pp. 1821–1852, Nov. 2019, doi: 10.1007/s10618-019-00647-x.

[65] H. Sedghamiz, "BioSigKit: A MATLAB toolbox and interface for analysis of BioSignals," *J. Open Source Softw.*, vol. 3, no. 30, p. 671, Oct. 2018, doi: 10.21105/joss.00671.

[66] M. Vollmer, "HRVTool—An open-source MATLAB toolbox for analyzing heart rate variability," in *Proc. Comput. Cardiology Conf. (CinC)*, Dec. 2019, pp. 1–4, doi: 10.22489/cinc.2019.032.

[67] M. P. Tarvainen, J.-P. Niskanen, J. A. Lipponen, P. O. Ranta-aho, and P. A. Karjalainen, "Kubios HRV—Heart rate variability analysis software," *Comput. Methods Programs Biomed.*, vol. 113, no. 1, pp. 210–220, Jan. 2014, doi: 10.1016/j.cmpb.2013.07.024.

[68] T. Mar, S. Zaunseder, J. P. Martínez, M. Llamedo, and R. Poll, "Optimization of ECG classification by means of feature selection," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 8, pp. 2168–2177, Aug. 2011, doi: 10.1109/tbme.2011.2113395.

[69] R. K. Samala, H.-P. Chan, L. Hadjiiski, and S. Koneru, "Hazards of data leakage in machine learning: A study on classification of breast cancer using deep neural networks," in *Proc. SPIE*, vol. 11314, Mar. 2020, Art. no. 1131416, doi: 10.1117/12.2549313.

[70] A. Pérez, P. Larrañaga, and I. Inza, "Bayesian classifiers based on kernel density estimation: Flexible classifiers," *Int. J. Approx. Reasoning*, vol. 50, no. 2, pp. 341–362, Feb. 2009, doi: 10.1016/j.ijar.2008.08.008.

[71] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, Dec. 2020, doi: 10.1186/s12864-019-6413-7.

[72] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/tkde.2008.239.

[73] R. Zhu, Z. Wang, Z. Ma, G. Wang, and J.-H. Xue, "LRID: A new metric of multi-class imbalance degree based on likelihood-ratio test," *Pattern Recognit. Lett.*, vol. 116, pp. 36–42, Dec. 2018, doi: 10.1016/j.patrec.2018.09.012.

[74] S. McGee, "Simplifying likelihood ratios," *J. Gen. Internal Med.*, vol. 17, no. 8, pp. 647–650, Aug. 2002, doi: 10.1046/j.1525-1497.2002.10750.x.

[75] J. Cohen, *Statistical Power Analysis for the Behavioural Sciences*. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 1988.

[76] D. Segebarth, M. Griebel, N. Stein, C. R. von Collenberg, C. Martin, D. Fiedler, L. B. Comeras, A. Sah, V. Schoeffler, T. Luffe, A. Durr, R. Gupta, M. Sasi, C. Lillesaar, M. D. Lange, R. O. Tasan, N. Singewald, H.-C. Pape, C. M. Flath, and R. Blum, "On the objectivity, reliability, and validity of deep learning enabled bioimage analyses," *eLife*, vol. 9, Oct. 2020, Art. no. e59780, doi: 10.7554/elife.59780.

[77] T. Finseth, M. C. Dorneich, N. Keren, W. D. Franke, and S. B. Vardeman, "Manipulating stress responses during spaceflight training with virtual stressors," *Appl. Sci.*, vol. 12, no. 5, p. 2289, Feb. 2022, doi: 10.3390/app12052289.

[78] M. Fallahi, R. Heidarimoghadam, M. Motamedzade, and M. Farhadian, "Psycho physiological and subjective responses to mental workload levels during N-back task," *J. Ergonom.*, vol. 6, no. 6, pp. 1–7, 2016, doi: 10.4172/2165-7556.1000181.

[79] G. Boateng and D. Kotz, "StressAware: An app for real-time stress monitoring on the amulet wearable platform," in *Proc. IEEE MIT Undergraduate Res. Technol. Conf. (URTC)*, Nov. 2016, pp. 1–4, doi: 10.1109/urtc.2016.8284068.

[80] I. Bichindaritz, C. Breen, E. Cole, N. Keshan, and P. Parimi, "Feature selection and machine learning based multilevel stress detection from ECG signals," in *Proc. Int. Conf. Innov. Med. Healthcare*, 2017, pp. 202–213, doi: 10.1007/978-3-319-59397-5_22.

[81] Y. S. Can, N. Chalabianloo, D. Ekiz, and C. Ersoy, "Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study," *Sensors*, vol. 19, no. 8, p. 1849, Apr. 2019, doi: 10.3390/s19081849.

[82] J. Salkevicius, R. Damasevicius, R. Maskeliunas, and I. Laukiene, "Anxiety level recognition for virtual reality therapy system using physiological signals," *Electronics*, vol. 8, no. 9, p. 1039, Sep. 2019, doi: 10.3390/electronics8091039.

[83] N. Keshan, P. V. Parimi, and I. Bichindaritz, "Machine learning for stress detection from ECG signals in automobile drivers," in *Proc. IEEE Int. Conf. Big Data*, Oct. 2015, pp. 2661–2669, doi: 10.1109/bigdata.2015.7364066.

[84] A. Arsalan, M. Majid, A. R. Butt, and S. M. Anwar, "Classification of perceived mental stress using a commercially available EEG headband," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 6, pp. 2257–2264, Nov. 2019, doi: 10.1109/jbhi.2019.2926407.

[85] M. Fredrikson, J. A. Blumenthal, D. D. Evans, A. Sherwood, and K. C. Light, "Cardiovascular responses in the laboratory and in the natural environment: Is blood pressure reactivity to laboratory-induced mental stress related to ambulatory blood pressure during everyday life?" *J. Psychosom. Res.*, vol. 33, no. 6, pp. 753–762, 1989, doi: 10.1016/0022-3999(89)90091-3.

[86] A. M. Langager, B. E. Hammerberg, D. L. Rotella, and H. M. Stauss, "Very low-frequency blood pressure variability depends on voltage-gated L-type Ca^{2+} channels in conscious rats," *Amer. J. Physiol.-Heart Circulatory Physiol.*, vol. 292, no. 3, pp. 1321–1327, Mar. 2007, doi: 10.1152/ajpheart.00874.2006.

[87] H. F. Posada-Quintero, J. P. Florian, A. D. Orjuela-Canon, T. Aljama-Corrales, S. Charleston-Villalobos, and K. H. Chon, "Power spectral density analysis of electrodermal activity for sympathetic function assessment," *Ann. Biomed. Eng.*, vol. 44, no. 10, pp. 3124–3135, Oct. 2016, doi: 10.1007/s10439-016-1606-6.

[88] C. Davatzikos, K. Ruparel, Y. Fan, D. G. Shen, M. Acharyya, J. W. Loughead, R. C. Gur, and D. D. Langleben, "Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection," *NeuroImage*, vol. 28, no. 3, pp. 663–668, Nov. 2005, doi: 10.1016/j.neuroimage.2005.08.009.

[89] G. G. Berntson, J. T. Cacioppo, P. F. Binkley, B. N. Uchino, K. S. Quigley, and A. Fieldstone, "Autonomic cardiac control. III. Psychological stress and cardiac response in autonomic space as revealed by pharmacological blockades," *Psychophysiology*, vol. 31, no. 6, pp. 599–608, Nov. 1994, doi: 10.1111/j.1469-8986.1994.tb02352.x.

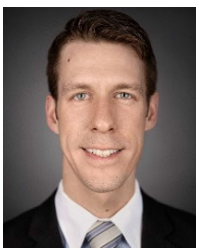
[90] K. Chrysostomou, S. Y. Chen, and X. Liu, "Combining multiple classifiers for wrapper feature selection," *Int. J. Data Mining, Model. Manag.*, vol. 1, no. 1, p. 91, 2008, doi: 10.1504/ijdmmm.2008.022539.

[91] B. S. Tegegne, T. Man, A. M. Van Roon, H. Riese, and H. Snieder, "Determinants of heart rate variability in the general population: The lifelines cohort study," *Heart Rhythm*, vol. 15, no. 10, pp. 1552–1558, Oct. 2018, doi: 10.1016/j.hrthm.2018.05.006.

[92] H. Chen, X. Liu, X. Li, H. Shi, and G. Zhao, "Analyze spontaneous gestures for emotional stress state recognition: A micro-gesture dataset and analysis with deep learning," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–8, doi: 10.1109/FG.2019.8756513.

[93] N. T. M. Chen, P. J. F. Clarke, T. L. Watson, C. MacLeod, and A. J. Guastella, "Biased saccadic responses to emotional stimuli in anxiety: An antisaccade study," *PLoS ONE*, vol. 9, no. 2, Feb. 2014, Art. no. e86474, doi: 10.1371/journal.pone.0086474.

[94] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLoS ONE*, vol. 14, no. 11, Nov. 2019, Art. no. e0224365, doi: 10.1371/journal.pone.0224365.
[95] B. Wang, C. Zhao, Z. Wang, K.-A. Yang, X. Cheng, W. Liu, W. Yu, S. Lin, Y. Zhao, K. M. Cheung, H. Lin, H. Hojaiji, P. S. Weiss, M. N. Stojanovic, A. J. Tomiyama, A. M. Andrews, and S. Emaminejad, "Wearable aptamer-field-effect transistor sensing system for noninvasive cortisol monitoring," *Sci. Adv.*, vol. 8, no. 1, Jan. 2022, Art. no. eabk0967, doi: 10.1126/sciadv.abk0967.
[96] E. A. Shirtcliff, R. L. Buck, M. J. Laughlin, T. Hart, C. R. Cole, and P. D. Slowey, "Salivary cortisol results obtainable within minutes of sample collection correspond with traditional immunoassays," *Clin. Therapeutics*, vol. 37, no. 3, pp. 505–514, Mar. 2015, doi: 10.1016/j.clinthera.2015.02.014.
[97] B. M. Kudielka and S. Wust, "Human models in acute and chronic stress: Assessing determinants of individual hypothalamus-pituitary-adrenal axis activity and reactivity," *Stress*, vol. 13, no. 1, pp. 1–14, Jan. 2010, doi: 10.3109/10253890902874913.
[98] J. Campbell and U. Ehlert, "Acute psychosocial stress: Does the emotional stress response correspond with physiological responses?" *Psychoneuroendocrinology*, vol. 37, no. 8, pp. 1111–1134, Aug. 2012, doi: 10.1016/j.psyneuen.2011.12.010.
[99] B. N. Uchino, J. Holt-Lunstad, L. E. Bloor, and R. A. Campo, "Aging and cardiovascular reactivity to stress: Longitudinal evidence for changes in stress reactivity.," *Psychol. Aging*, vol. 20, no. 1, pp. 134–143, 2005, doi: 10.1037/0882-7974.20.1.134.
[100] L.-L. Chen, Y. Zhao, P.-F. Ye, J. Zhang, and J.-Z. Zou, "Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers," *Exp. Syst. Appl.*, vol. 85, pp. 279–291, Nov. 2017, doi: 10.1016/j.eswa.2017.01.040.
[101] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, and F. Hormozdiari, "Underspecification presents challenges for credibility in modern machine learning," 2020, *arXiv:2011.03395*.
[102] S. Marionoryad and C. Busso, "The cost of dichotomizing continuous labels for binary classification problems: Deriving a Bayesian-optimal classifier," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 119–130, Jan./Mar. 2017, doi: 10.1109/taffc.2015.2508454.



TOR T. FINSETH (Member, IEEE) received the B.S., M.Eng., and Ph.D. degrees in aerospace engineering from Iowa State University, Ames, IA, USA, in 2012, 2016, and 2021, respectively.

His research interests include the biomarkers for the human stress response, psychosocial stress, virtual reality, acute stress detection, adaptive systems, and cognition.

MICHAEL C. DORNEICH (Senior Member, IEEE) received the Ph.D. degree in industrial engineering in the human factors program from the University of Illinois Urbana–Champaign, Champaign, IL, USA, in 1999.

He is currently a Professor with the Industrial and Manufacturing Systems Engineering Department, Iowa State University, Ames, IA, USA. His research interest includes creating adaptive joint human–machine systems that enable people to be effective in the complex environments.

STEPHEN VARDEMAN received the B.S. and M.S. degrees in mathematics from Iowa State University, Ames, IA, USA, in 1971 and 1973, respectively, and the Ph.D. degree in statistics from Michigan State University, East Lansing, MI, USA, in 1975.

From 1975 to 1981, he was an Assistant Professor of statistics with Purdue University. From 1981 to 2021, he was a Faculty Member with the Statistics Department and the Industrial and Manufacturing Systems Department, Iowa State University, Ames, IA, where he has been an Emeritus University Professor of statistics and industrial and manufacturing systems, since 2021.

Prof. Vardeman is a fellow of the American Statistical Association. He was the winner of the 1994 ASEE and the Meriam/Wiley Distinguished Author Award for a new engineering textbook. He was an Editor of the *Technometrics* (ASA) journal, from 1993 to 1995.

NIR KEREN received the B.S. and M.S. degrees from Ben Gurion University, Israel, in 1990 and 1998, respectively, and the Ph.D. degree from Texas A&M University, in 2003. He is currently an Associate Professor of occupational safety with Iowa State University and the Director of the Occupational Safety Program, NIOSH Heartland Education and Research Center for Occupational Health and Safety. He is a Graduate Faculty with the Human–Computer Interaction Program, Iowa State University, where he is also a Faculty Fellow with the Virtual Reality Applications Center and the Institute of Transportation. His research interests include using virtual reality environments to study mission-critical occupation employees' performance under stress and the capacity of VR as a tool in enhancing design performance and increasing training transfer.

WARREN D. FRANKE received the Ph.D. degree in applied exercise physiology from Virginia Tech, Blacksburg, VA, USA, in 1991.

He is currently a Professor with the Department of Kinesiology, Iowa State University, Ames, IA, USA. His research interests include mechanisms underlying cardiovascular disease in emergency responders, using virtual reality environments to assess decision-making and cardiovascular responses to occupationally relevant stressful situations, exercise in older adults, and cardiovascular exercise physiology.