CLACIR: Dataset for Cognitive Load and Canine Intervention Recognition from Physiological Signals

Compiled on 2025/12/28 at 16:41:14

Walker S. Arce

University of Nebraska-Lincoln, Department of Electrical and Computer Engineering, Lincoln, 68588, United States

E-mail: wsarcera@gmail.com

Jeffrey R. Stevens

University of Nebraska-Lincoln, Department of Psychology, Center for Brain, Biology & Behavior, Lincoln, 68588, United States

E-mail: jeffrey.r.stevens@gmail.com

Abstract. Emotionally intelligent machines must differentiate human emotional states in the face of subtle differences and variation in states across individuals. Large datasets are needed to train models how to classify different emotional states. Here, we introduce the Cognitive Load and Canine Intervention Recognition (CLACIR) dataset, which consists of 95.8 hours of physiological responses recorded using an Empatica E4 to cognitive load and subsequent intervention from 140 participants. Physiological data consist of 3-axis accelerometry, electrodermal activity, and photoplethysmography. To validate this dataset, we employed machine learning algorithms to classify different states in the dataset with a stratified group 10-fold cross validation. We found that Linear Discriminant Analysis best distinguished between intervention classes with an accuracy of 83.5% ± 3.7%. Accelerometry measures provided the most important features for classification, with blood volume pulse measures coming in second. The CLACIR dataset performs similarly to other smaller datasets but provides a large sampling of participants with a well-defined protocol that induces a commonly investigated mental state of cognitive load and performs well with a variety of machine learning models.

Keywords: affective computing, machine learning, human-animal interaction, generalizability

Submitted to: *Machine Learning: Health*

Introduction

Human emotion and affect are physiological and behavioral reactions to changes in our environment [1]. Some affective states may be easy to label (e.g., anger) but others may be more difficult to pin down (e.g., stress, anxiety). What's more, individuals vary greatly in how they experience affective states, with some experiencing intense emotions and others experiencing more muted emotions. The field of affective computing develops machines and algorithms that can detect human affective states using measures of physiological responses such as heart rate, breathing rate, skin conductance, and muscle activation [1]. These physiological data are input into algorithms that attempt to classify the affective state experienced when the data were collected. Through affective computing, machines can infer the internal state of human emotion, potentially leading to improved monitoring of clinical populations during treatment [2, 3, 4, 5], monitoring player stress in simulated driving or airport environments [6, 7, 8], or to monitor stress levels in real-life environments to develop strategies to reduce it [9, 10].

Accurately classifying human affective state requires large data sets due to the variability in emotions experienced within and between people. But the number of large-scale datasets for physiological signals is limited. Collecting physiological data can be difficult because participants must wear recording equipment either during their daily life or in a laboratory study [11, 12, 13, 14, 15, 16, 17, 18]. The human body generates physiological signals through chemical and electrical interactions with various organs, which are largely involuntary and not easily controlled by an individual [14]. Consequently, targeted emotional states must either be induced in participants within a laboratory setting or annotated through surveys from naturalistic settings. Certain cognitive tasks can be used to reliably induce physiological states. For instance, researchers can put participants into a state of *cognitive load*—where cognitive limits of processing information are strained—by engaging long-term memory, working memory, and attentional control [19]. Physiological measures show that cognitive load can cause stress-like responses in the autonomic nervous system, including changes in heart rate, heart rate variability, and galvanic skin response [20]. However, previous work has found ways to discriminate cognitive load from stress. Notably, Setz et al. 2009 found that 33 participants undergoing cognitive load induction through a social-evaluative threat could have their cognitive load and stress states differentiated with an 82.8% accuracy using a wristworn EDA device [21]. Additionally, Markova, Ganchev, and Kalinkov 2019 collected physiological data from 62 participants using the Shimmer3 system to predict concentration, i.e. cognitive load, with an accuracy of 74.2% [22].

To improve the state-of-the-art, we introduce the Cognitive Load and Canine Intervention Recognition (CLACIR) dataset, which provides multimodal physiological data from a variety of mental states including active cognitive load, control intervention, active intervention, and post-intervention cognitive load. The dataset comes from a study of how human-animal interactions influence mood, stress, anxiety, and cognition [15].

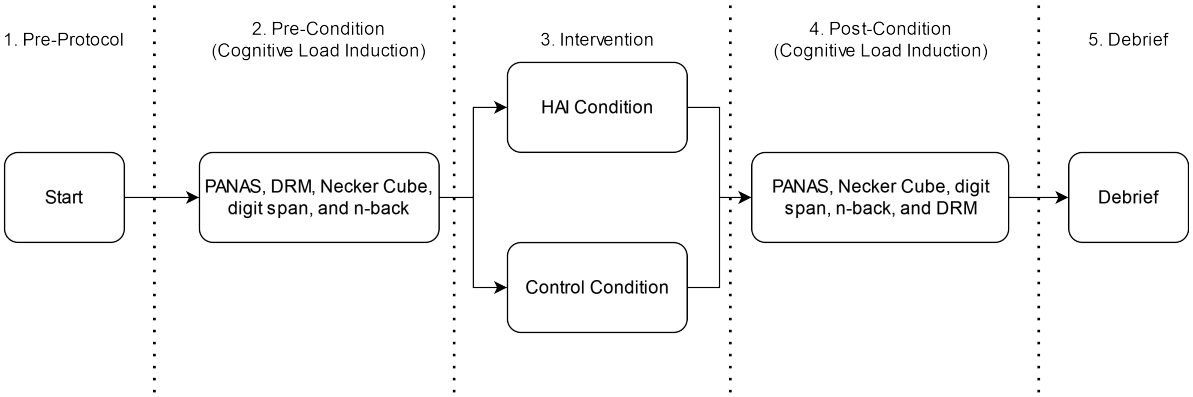

Figure 1. The experimental protocol to induce cognitive load states and provide intervention from the same using a control and human-animal interaction (HAI) conditions.

Our study tested 140 participants in three experimental phases: pre-condition, intervention, and post-condition (Figure 1). The pre-condition phases tested participant long-term memory, attentional control, and working memory. After completing these cognitive tasks, participants entered the intervention condition in which they either interacted with a dog (active intervention) or completed a repetitive task (control intervention). After the intervention, participants completed the same cognitive tasks. Thayer and Stevens, 2022 analyzed the cognitive and mood data but not the physiological data collected during the study. Physiological data included 95.8 hours of data on blood volume pulse, electrodermal activity, temperature, and movement from 140 participants.

In this preregistered [23] study, we sought to validate the CLACIR dataset by using machine learning algorithms to classify different affective states. Our first aim was to determine how accurately machine learning models can differentiate between stress intervention exposure. Our second aim was to determine which machine learning model best classified the stress states. Our third aim was to determine if the accelerometer acted as a naive discriminator between the two stress intervention states in our dataset. That is, the active intervention involved petting a dog, which could have resulted in more movement and thus specific accelerometer signals that would be easy to distinguish from the control intervention. Therefore, we wanted to determine if the other physiological measures could discriminate these conditions. The final aim was to determine which features of the input data were most important in predicting the stress intervention state.

Methods

Data Collection

This dataset was collected as a series of two experiments to investigate the effect of human-animal interactions (HAIs) on affect and cognitive ability of the participants [15].

The first experiment was conducted from September to November 2018 and recruited 73 participants, 13 of which identified as male (17.8%) and 60 identified as female (82.2%), with a mean age of 19.2 (SD = 1.4) years old. The second experiment was conducted from November 2018 to April 2019 and recruited 83 participants, 17 of which identified as male (20.5%) and 66 identified as female (79.5%), with a mean age of 19.9 (SD = 1.8) years old. All participants were recruited from the University of Nebraska-Lincoln's psychology subject pool and did not have a physical or emotional aversion to dogs. For 16 participants, the Empatica E4 did not properly collect physiological data, so they were excluded from this study. Additionally, due to a hardware issue with the Empatica E4, the temperature data was not reliable (reporting over 200 °F skin temperature), so it was excluded from this study.

To collect this data, an experimental protocol was designed that induced a state of cognitive load on the participant, then applied an intervention condition, and then applied another cognitive load condition (Figure 1). The cognitive load induction had four separate cognitive tasks: Deese-Roedinger-McDermott (DRM) to test long term memory [24, 25], the Necker cube pattern control to test attentional control [26, 27, 28], the backwards digit span to test working memory [29], and the n-back to test working memory [30, 31]. These assessments were presented and responses collected using PsychoPy version 1.90.2 [32] running on a computer's 16-inch monitor with only the participant and researcher present.

After the cognitive load state was induced in the participant, they moved on to the intervention condition. Each participant was pseudo randomly assigned to either a human-animal interaction (HAI) condition or a control condition. For the HAI condition, the participant interacted with the second author's dog, who was a 65-pound, neutered, male Catahoula leopard mix that was Canine Good Citizen certified prior to the study. For the control condition, the participant was alone in the room and provided a sheet of paper with a full page of Latin text and instructed to circle every "e" and "f". The participant was informed that the task would not be graded and there is no penalty for wrong answers. The intervention condition was performed for three minutes in both cases. After the intervention condition, the participant repeated the cognitive load induction activity.

To measure the participant's affect, the Positive and Negative Affect Schedule (PANAS) [33] was administered before each cognitive load induction. The second experiment also added measures of anxiety, including present and general feeling of anxiety with the State and Trait Anxiety Inventory (STAI) from [34] and their present feelings of anxiety using the Anxiety Visual Analogue Scale (AVAS) from [35]. Due to the new measures, the cognitive load induction tasks were modified to perform the AVAS first, then STAI, and then PANAS.

Physiological Data Collection

To measure the physiological responses of the participants, an Empatica E4 biosensor was affixed to their left wrist. This wrist-worn biosensor has been used to generate data in many studies [14, 13], is regulatory compliant and validated [17], and is the top rated device for patient populations served by the behavioral clinics [36]. This biosensor has four sensors: 64 Hz blood volume pulse (BVP) using photoplethysmography (PPG), 4 Hz electrodermal activity (EDA), 4 Hz infrared thermopile, and a 32 Hz 3-axis accelerometer measuring accelerations in its inertial frame of reference [14]. When transitioning between states of the protocol (start → cognitive load induction → intervention → cognitive load induction → debrief) researchers pressed the button on the biosensor to place a marker in the datastream. This marker labels each data point with its associated protocol stage during post-processing. Figure 2 provides a schematic representation of the hierarchical structure of the dataset.

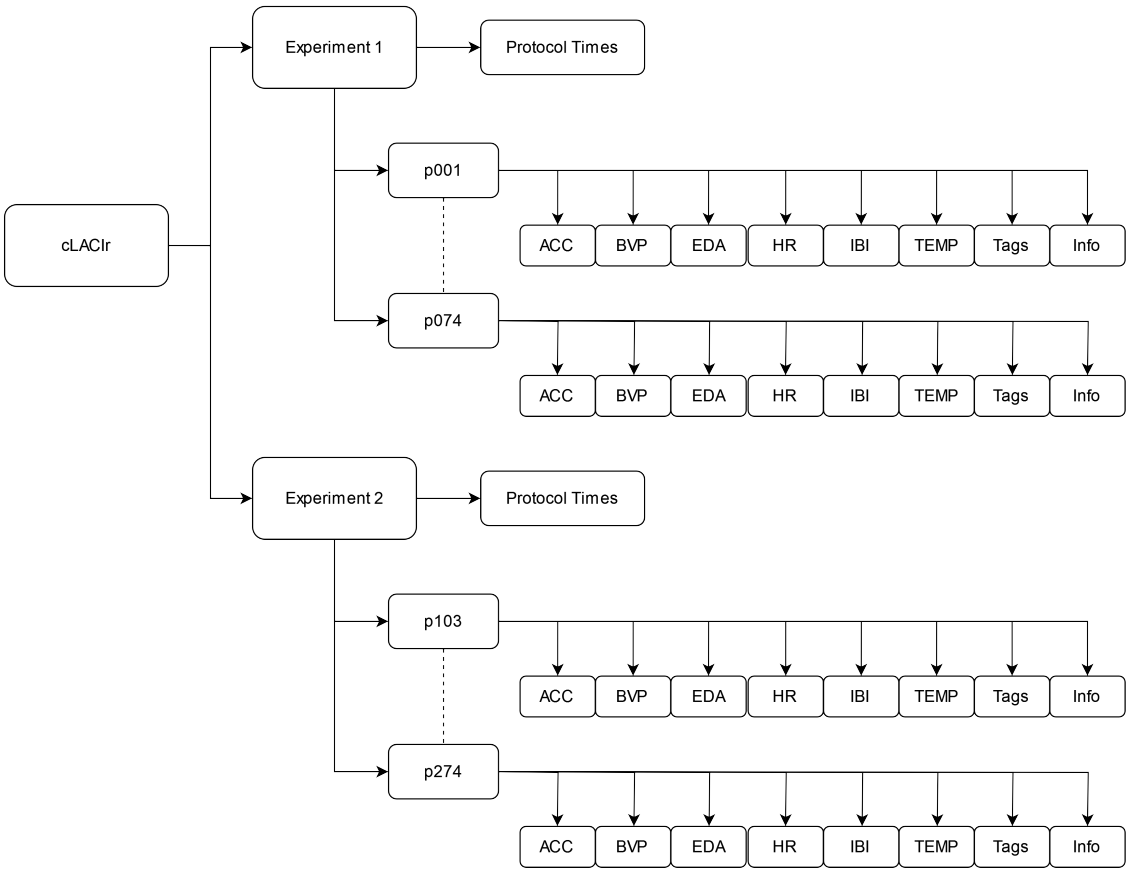

Figure 2. The hierarchical structure of the CLACIR dataset. The dataset consists of two experiements with 73 and 83 participants respectively. For each participant, the dataset includes information on accelerometry (ACC), blood volume pulse (BVP), electrodermal activity (EDA), heart rate (HR), inter-heartbeat interval (IBI), body temperature (TEMP), and tags/markers signaling the experimental phase.

The E4 has multiple methods to compensate for common issues with in-situ physiological signal recording. Measurement error in the PPG signal due to skin color

and changing light intensity is compensated dynamically by the device [14]. Motion artifacts are removed from the captured signals and heart rate variability artifacts are removed as well [14]. Having this preprocessing be performed on the data before it is reported to the collection software reduces the burden on preprocessing for this experiment. Due to this, minimal preprocessing is needed before training computational intelligence models.

Each physiological signal from the Empatica E4 encodes either physical movement of the participant (accelerometer) or their autonomic nervous system responses (PPG, EDA, thermopile). By tracking the movement of the participant's hand, specific movements that occur during emotional states can be assessed, which is the basis of physical activity recognition [37]. The separability of movements from different stimuli will likely be highly variable between participants and may be highly correlated to the level of reaction [38].

Autonomic nervous system responses are highly correlated with internal emotional state [39] and are the basis of emotion recognition in affective computing systems [40]. The most common physiological signal used for this task is the EDA signal, which is collected by measuring the resistance between two contacts pressed against the participant's skin [41, 6]. EDA data can be decomposed into three constituent signals: the skin conductance response (SCR), the skin conductance level (SCL), and noise [42]. The noise will be ignored. The SCR is the fast-changing response to events, constituting a phasic component of the EDA signal, and can be used to approximate sympathetic nervous system activity [41, 43]. The SCL is a slower changing, tonic component of the EDA signal that acts as the baseline conductance of the skin. These two components indicate how strongly an emotion was felt and can also be used to differentiate between emotional states, for instance, the SCL and SCR will go up when someone is feeling anxiety and only the SCL will go up if someone is feeling happiness [39].

Peripheral skin temperature is a common measure of stress and can be used to distinguish between emotional states [44, 45, 46]. Generally, the theory of blood flow constriction due to stress causing a decrease in skin temperature and blood flow dilation due to happiness causing an increase in skin temperature is used [46]. But there are conflicting reports of skin temperature response to emotional states, where few report the temperature of the wrist or hand and instead focus on either monitoring facial temperature with a thermal camera [45] or will place their temperature probe in an area that is free from thermal environmental disturbances, such as under the armpit [46]. For studies that place the temperature probe on the hand [44], a decrease in skin temperature is observed during happy states and an increase in skin temperature is observed for anger. Due to unusually high recorded values (< 200 °F), the observations from the temperature probe on the hand will not be considered.

Finally, PPG is an indirect method of measuring the cardiac cycle and can be compared with the electrocardiogram (ECG), which directly measures the electrical activity of the heart [47]. PPG is performed by measuring the change in skin coloration as blood travels through the capillaries of the skin, which is directly correlated to the

blood being pumped by the heart (blood volume pulse—BVP). From BVP, the Empatica system automatically calculates heart rate (HR) in beats per minute and the inter-beat interval of heartbeats (IBI). In the case of someone feeling anxiety and happiness, their heart rate will go up [39].

Though the cognitive and affective data were reported in [15], the physiological data were not analyzed in the original study due to low sampling rate of the heart rate signal [48]. The physiological data have been analyzed using machine learning in the first author's master's thesis [49].

Data Preprocessing

To process the dataset, handcrafted features were extracted from overlapping 60-second windows of data. These features consist of statistical information generated using the FLIRT toolbox [50]. The EDA signal is decomposed into its constituent signals (SCL, SCR, and noise) using cvxEDA [42]. We kept the original signal, SCL, and SCR but discarded the noise component. The other signals (PPG, accelerometry, and skin temperature) were processed in their original form using FLIRT. The calculated statistics for each signal window were the mean, standard deviation, minimum, maximum, dynamic range, sum, energy, skewness, kurtosis, peak count, RMS, line integral, number of samples above and below mean, number of sign changes, 25%-75% interquartile range, 5%-95% interquartile range, 5-th percentile, and 95-th percentile. Additionally, the entropy, perumutation entropy, and singular value decomposition entropy were calculated. We calculated statistics for each signal broken into 5-second windows, which is a common approach [13, 2]. Each participant's data was tagged with their participant number, allowing the identity of the data to be tracked during training.

The data windows were labeled using the protocol steps and the dataset was narrowed down to the intervention stage. The participants in the HAI condition were labeled with a one and participants in the control condition were labeled with a zero, constructing a binary classification task.

Machine Learning Approaches

The dataset will be evaluated using a standard selection of common machine learning models. These models are Linear Discriminant Analysis (LDA), Decision Tree Classifier (DTC), Random Forest Classifier (RFC), and the K-Nearest Neighbors algorithm (KNC), all of which are implemented in SciKit-Learn [51]. An LDA model uses a linear combination of features to separate two or more classes, where predictions are made using Bayes' rule where the selected class maximizes the posterior probability [52]. These models are also closely related to analysis of variance (ANOVA), but the LDA uses continuous independent variables and categorical dependent variables [53]. The DTC model uses simple decision rules, such as feature a is less than feature b infers class a, which makes predictions very explainable [54]. The RFC is a meta-estimator

that uses an ensemble of DTCs to predict classes based on the input data by taking a majority vote from the DTC ensemble [55].

To train the machine learning models, a stratified group 10-fold approach was used [51]. Using the identity labels, dataset features, and dataset labels, 10 splits of the data were generated, where each split contains a different group of identities and the number of samples from each class was as balanced as possible. By doing this, not only will the test set contain unique identities from the train set, a requirement for generalization performance, but the choice of identity for testing can also be evaluated. Doing this, a mean and standard deviation for each evaluation metric was generated. If there are ideal and non-ideal choices for test participant, then this will be observed as a high standard deviation in the final metrics. Finally, a PANAS threshold was generated to investigate how the PANAS data aligns with affective computing capability. The PANAS threshold is used to filter out participants based on the difference between their pre- and post-PANAS positive affect scores, so filtering out participants below -0.5 would remove participants whose positive affect decreased by 0.5 from pre-condition to post-condition. By varying this from -1 to 1, we can filter out participants who had an increasingly positive reaction to the data collection and see if the physiological data reflects this.

Figures of Merit

While the training of ML models can be simplified to a few function calls using highly abstracted software libraries, the methods to compare trained models in a fair way is still an active question. The most common method of comparison is accuracy, or more properly, top-1 accuracy. This metric quantifies how often the highest probability class is the target class. There is also top-k accuracy, which looks at an increasing number of output probabilities and, if the target class is in that list of probabilities, then it counts as correct [56]. For binary classification, there is only top-1 accuracy.

The next most common metrics are recall- and precision-based metrics. Recall as a metric quantifies how often the model correctly chooses a class versus the number of times that it should have chosen that class, while precision quantifies how often a class is chosen correctly versus all the times that class is chosen [57, 56]. These two metrics can be combined to create derivative metrics, such as the F1 score, which is the harmonic mean of precision and recall [56] and can give insight into how well a model balances those two metrics. At this level, these mentioned metrics are common across all ML analyses.

Computer vision has utilized many robust metrics using receiver operator characteristic (ROC) curves to quantify the diagnostic ability of a binary classifier at varying threshold levels on the probability outputs [57, 58, 56]. For these curves, the true positive rate (TPR) is plotted on the y-axis and the false positive rate (FPR) is plotted on the x-axis for all values of the threshold from 0 to 1. An optimal relationship between TPR and FPR occurs when the TPR is very high and the FPR is very low,

meaning that the curve approaches the upper left corner of the plot. The area under the curve (AUC) of the ROC curve is a way of capturing the curve in a single number, where the closer it is to one, the closer the ROC curve is to the upper left of the plot area. Setting thresholds on the FPR can be useful for establishing a confidence level in the performance of a model. If the FPR is set to 1%, then a corresponding TPR of 95% would indicate that for every 1 false positive (FP) prediction, 95 true positive (TP) predictions would be accomplished on average. This is commonly extended to two or three thresholds, namely: 1%, 5%, and 10%. Equally important to mention is the equal error rate (EER), which is the ratio of the FPR to the false negative rate (FNR) by drawing a line from the top left corner of the ROC curve to the bottom right corner and finding the intersecting values [59]. In this case, a lower value indicates that the classifier is performing better, and this metric is important when the risk of false positive or false negative identification is equally harmful.

While ROC metrics such as AUROC are used in affective computing research, it is not widespread and isn't accompanied by other ROC metrics, such as TPR @ x% FPR, which would provide heightened insight into the confidence of a classifier. Other metrics are being discussed for affective computing, namely the area under the precision recall curve (AUPRC) and Cohen's kappa coefficient [60]. AUPRC is constructed similarly to the AUROC, but precision is plotted on the y-axis and recall is plotted on the x-axis for all values of the threshold, when both metrics are high then the curve will increase towards the upper right corner of the plot area. The larger the area under the curve, the greater the precision and recall at all thresholds. This metric is also called average precision (AP) and in multi-class problems, the mean AP (mAP) is the average AUPRC across all classes. Cohen's kappa coefficient quantifies the inter-rater reliability, which is the degree of agreement between independent observers of the same phenomena, which can be considered a more robust accuracy measure because it takes the probability of random chance agreements into account [61]. The concern of not using descriptive metrics is related to the way that AC tasks are constructed. In most cases, the classification task of emotion vs. neutral has a data distribution that leans towards neutral [13, 60]. Consequently, a model trained on this imbalanced data, i.e., 80% neutral / 20% emotion, then predicting zero (neutral) for every sample will generate a top-1 accuracy score of 80% automatically [60]. Increasing in popularity, balanced accuracy seeks to improve top-1 accuracy by using the TPR plus the true negative rate (TNR) divided by two, which accounts for some of the class imbalance problems.

The final model evaluation method to be used is permutation feature importance [62], which utilizes the trained model and the test set of data where one feature of the dataset is randomized to assess the impact on the output scores. The output score increasing or decreasing can indicate how the combination of your features contribute to your trained score. For the best models, the permutation feature importance will be calculated and compared.

Pre-registration and Data Availability

The analyses for this study were pre-registered at `https://https://osf.io/5k9ra/`. Data are available at `https://github.com/unl-cchil/clacir_dataset`.

Results

HAI vs. Control Intervention Classification

The first classification test of the dataset differentiated between the HAI condition and the control condition during the intervention stage of the data collection (Figure 3).

With the full dataset (PANAS threshold of -1), classification accuracies ranged between 75-83% depending on the model. The RFC and LDA perform the best of the four, with balanced accuracy staying above 80% after training. This indicates that using all of the physiological data, RFC and LDA can accurately classify the label as the HAI or control condition more than 80% of the time. Using a threshold on the difference between the pre- and post-PANAS survey positive affect, the accuracy is fairly constant with a slight bump to the balanced accuracy results when the PANAS threshold between 0.7-0.8. This indicates that removing participants with a decreasing PANAS positive affect during the study improves the ability to differentiate between classes. This contrasts with the other metrics, where AUPRC continues to drop as the PANAS threshold increases. The Kappa test shows some loss in agreement between raters, but the TPR @ 10% FPR shows a similar trend to the balanced accuracy, where it stays solid until the PANAS threshold is sufficiently high and it gets a boost. The models all show similar comparisons to one another across metrics (no model outperforms another in one metric that it underperforms in another), which is not entirely unexpected. Interestingly, the DTC model has a horrible TPR @ 10% FPR, which steadily increases as the PANAS threshold is tuned. At a -1 PANAS threshold, LDA performs the best with a balanced accuracy of 83.5 $\pm$ 3.7, a Kappa of 66.6 $\pm$ 7.5, an AUPRC of 92.5 $\pm$ 3.6, and a TPR @ 10% FPR of 77.9 $\pm$ 7.2. So, to answer our first research question, we can classify the intervention type with 83.5% $\pm$ 3.7% accuracy, and to answer our second research question, Linear Discriminant Analysis is the most accurate model, but the Random Forest Classifier is a close second.

HAI vs. Control Intervention Classification without Accelerometer

The second classification test of the dataset differentiated between HAI and control conditions during the intervention stage of the data collection, but without accelerometry (Figure 4).

Based on the theory that accelerometry acts as an clear differentiator between the HAI condition and the control condition, we expected that the performance would degrade. In the worst case, the performance could completely collapse to random chance because the other features do not meaningfully map onto the two states. Comparing

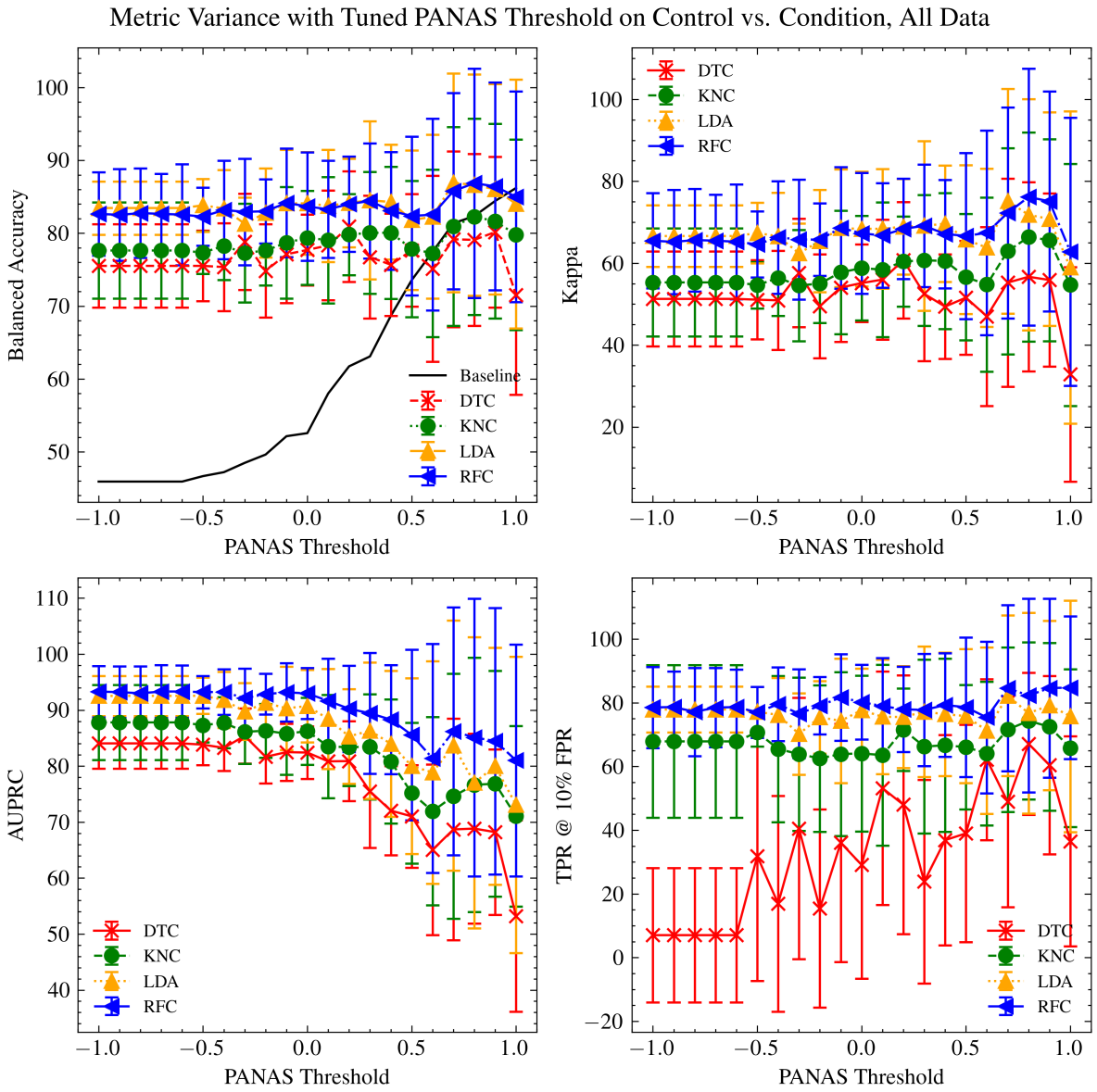

Figure 3. The effect of using various thresholds on the PANAS score to exclude participants from dataset and training to differentiate between intervention conditions using data from intervention stage. Using the difference between the pre- and post-PANAS positive affect survey values, the PANAS threshold is used to filter out participants that are below the given value. Panels represent different measures of accuracy, including balanced accuracy, Cohen's kappa coefficient, area under the precision recall curve (AUPRC), and true positive rate at 10% of false positive rate (TPR @ 10% FPR). Individual lines represent mean±standard deviation performance for decision tree classifier (DTC), k-nearest neighbors (KNN), linear discriminant analysis (LDA), and random forest classifier (RFC). For balanced accuracy, the black line represents the baseline value, which is the ratio of one class over another.

Figures 3 and 4, all metrics reduced in performance when acceleometry was removed. Similar trends emerged from varying the PANAS threshold value, but the KNC and DTC models performed worse overall without the accelerometry. This reduction in

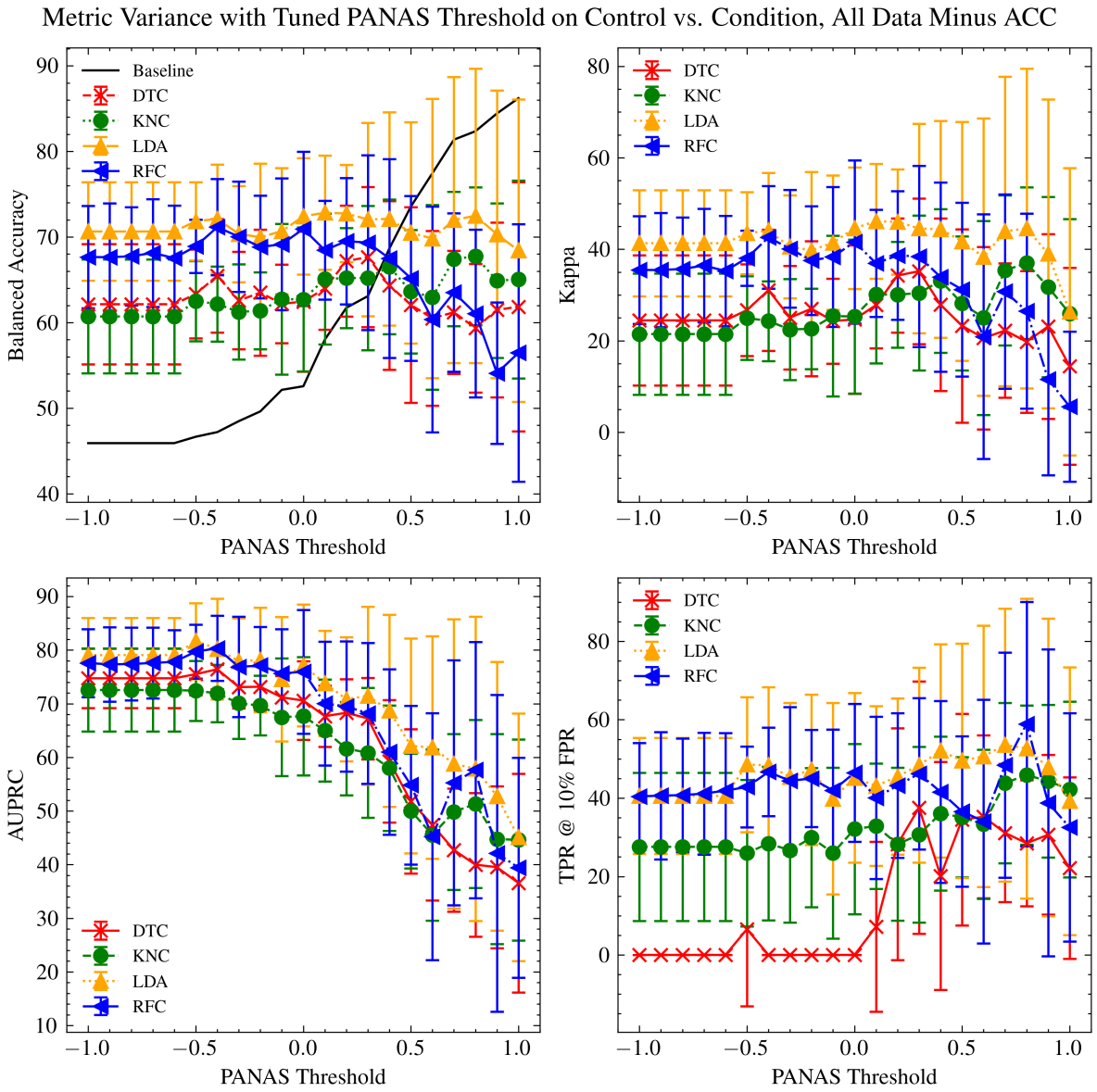

Figure 4. The effect of using various thresholds on the PANAS score to exclude participants from dataset and training to differentiate between intervention conditions using data from the intervention stage, but without accelerometry. Using the difference between the pre- and post-PANAS positive affect survey values, the PANAS threshold is used to filter out participants that are below the given value. Panels represent different measures of accuracy, including balanced accuracy, Cohen's kappa coefficient, area under the precision recall curve (AUPRC), and true positive rate at 10% of false positive rate (TPR @ 10% FPR). Individual lines represent mean±standard deviation performance for decision tree classifier (DTC), k-nearest neighbors (KNN), linear discriminant analysis (LDA), and random forest classifier (RFC). For balanced accuracy, the black line represents the baseline value, which is the ratio of one class over another.

performance indicates that accelerometry provides a meaningful impact to the predictive power of the model. For LDA with a -1 PANAS threshold, the balanced accuracy is

353 70.6 ± 5.8, Kappa is 41.3 ± 11.6, AUPRC is 79.1 ± 6.9, and TPR @ 10% FPR is 40.6 ±
354 14.8. To answer our third research question, removing accelerometry reduces the highest
355 accuracy rates by 12 percentage points but maintains the same best model type.

356 *Feature Importance*

357 The permutation feature importance was not calculated for the KNC model due to
358 computation time requirements. Each run of the permutation feature importance was
359 conducted ten times to ensure variability is accounted for. For the RFC models using all
360 variables (Figure 3) with PANAS threshold -1, the ten top features were ACC features.
361 For the LDA models using all variables with PANAS threshold -1, four of the top ten
362 features were ACC features, two were BVP, and four were HRV. For the LDA models
363 removing accelerometry variables (Figure 4) with PANAS threshold -1, two of the top
364 ten features were BVP and the remaining eight were HRV.

365 The ACC data stream results in the best performance, but this is likely due to the
366 distinct hand movement patterns observed in the HAI and control conditions. The
367 permutation feature importance for the RFC models were very low (<1%), which
368 indicates significant correlation between many of the features. This was explored by
369 utilizing a modified permutation feature importance where the randomized feature is
370 kept between runs. By doing this, the test set would go from the original data to
371 fully randomized. This should show the degree of correlation between features. If
372 the correlation is high, then many features can be randomized without fully collapsing
373 model performance because predictive information is being captured by many of the
374 non-randomized features as well.

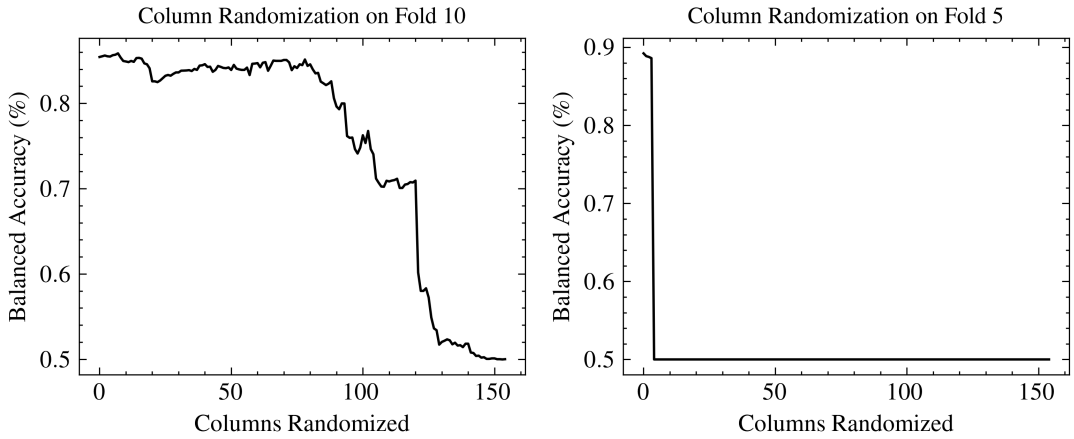

Figure 5. The balanced accuracy of an RFC model (left) and an LDA model (right) from Figure 4. As more features are randomized in the dataset, the original performance of the model degrades.

375 Figure 5 illustrates the stronger effects of feature correlation for RFC model (left
376 side) compared to LDA model (right side). The RFC model high accuracy levels until 90
377 features were randomized, explaining the low feature importance for RFC. On the other

hand, the LDA model instantly collapses after three features are randomized, suggesting these features were critical to the LDA model's classification.

Individual Data Stream Performance

It became clear that adding or removing different data streams (e.g., ACC, EDA, BVP) from the dataset impact the predictive power of the trained models. To further explore this, each data stream was isolated for learning to differentiate between HAI and control during the intervention stage.

When the dataset is cut down to only EDA (Figure 6), the overall performance is lower than when all variables are included (Figure 3), but LDA outperforms all other models. With a PANAS threshold of -1, LDA has a balanced accuracy of 64.6 ± 5.4, a Kappa of 29.0 ± 10.9, an AUPRC of 72.9 ± 7.0, and a TPR @ 10% FPR of 31.7 ± 15.3.

With only BVP in Figure 7, LDA clearly outperforms the other models, with RFC in second place. Compared to Figure 6, Figure 7 improved in performance, but there is still a reduction in overall performance compared to Figure 3. For the LDA model with a PANAS threshold of -1, balanced accuracy was 71.4 ± 5.3, Kappa was 42.9 ± 10.6, AUPRC was 80.9 ± 6.2, and TPR @ 10% FPR was 46.2 ± 12.0.

With only HRV in Figure 8 LDA outperformed the models with a balanced accuracy of 63.2 ± 7.3, a Kappa of 26.5 ± 14.7, an AUPRC of 70.3 ± 9.5, and a TPR @ 10% FPR of 25.6 ± 10.5.

When using only ACC in Figure 9, the model performance looks very similar to Figure 3. This indicates that the models trained on all data streams get a boost from the ACC data. The LDA model had a balanced accuracy of 85.3 ± 4.8, a Kappa of 70.2 ± 10.0, an AUPRC of 94.3 ± 3.3, and a TPR @ 10% FPR of 81.8 ± 9.2.

Discussion

We report the results of training various machine learning models with the Cognitive Load and Canine Intervention Recognition dataset, evaluating the performance of the stress intervention recognition using various figures of merit, and determining the most important features for inference. The best machine learning model for classifying the stress intervention class was the Linear Discriminant Analysis, which had a balanced accuracy of 83.5 ± 3.7.

When varying the PANAS threshold, the baseline performance increased because the participants who responded more positively to the experimental protocol interacted with the dog. While this led to an increase in the balanced accuracy, a decrease in the other metrics could be observed. If only balanced accuracy was reported, then this could bias the reported results. Some of the models showed an increase in all metrics with a PANAS threshold of 0.7, indicating that limiting participant data can be benefit model performance. Finally, the standard deviation increases as the PANAS threshold

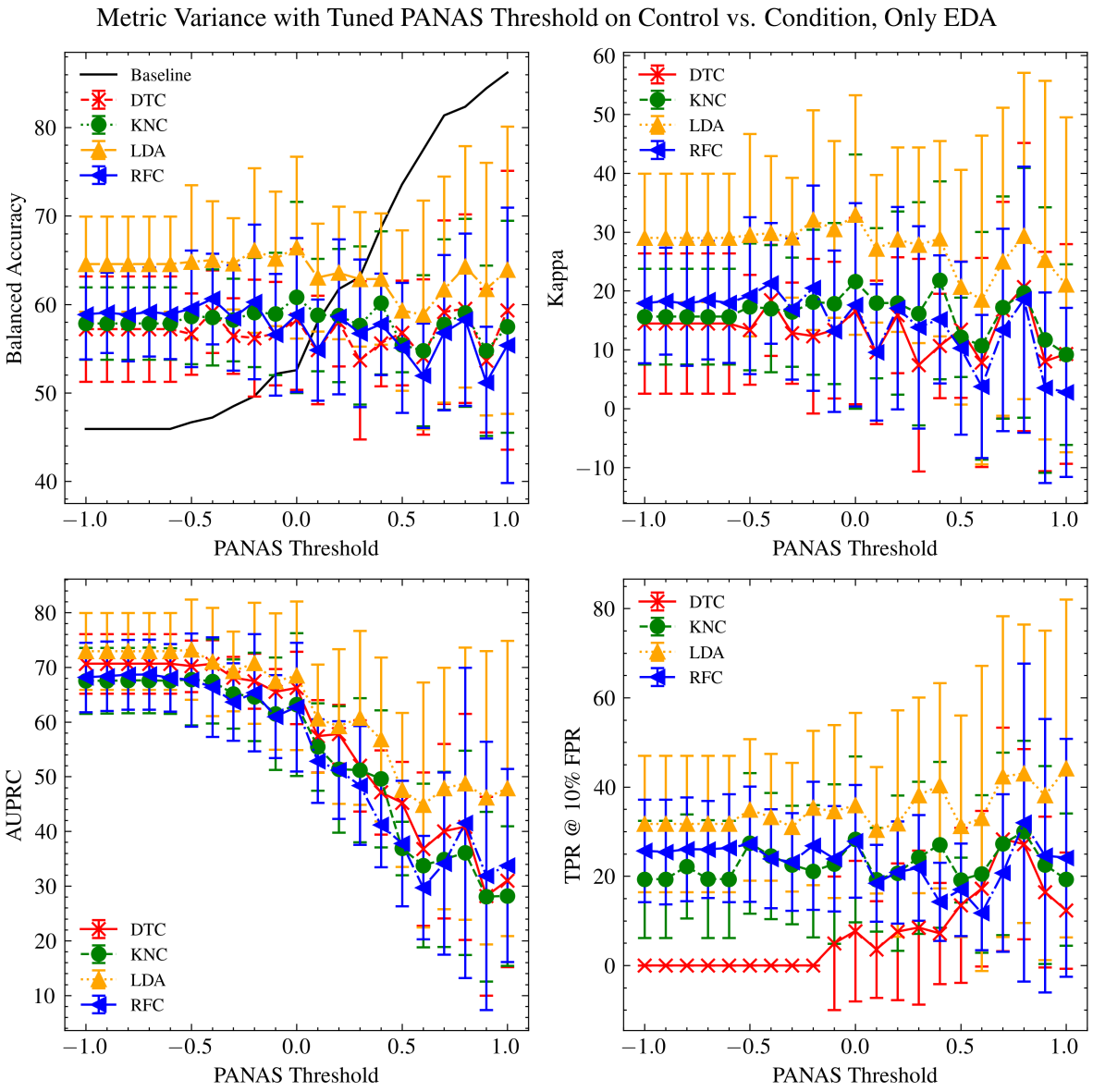

Figure 6. The effect of using various thresholds on the PANAS score to exclude participants from dataset and training to differentiate between intervention conditions using data from the intervention stage, using only EDA. Using the difference between the pre- and post-PANAS positive affect survey values, the PANAS threshold is used to filter out participants that are below the given value. Panels represent different measures of accuracy, including balanced accuracy, Cohen's kappa coefficient, area under the precision recall curve (AUPRC), and true positive rate at 10% of false positive rate (TPR @ 10% FPR). Individual lines represent mean±standard deviation performance for decision tree classifier (DTC), k-nearest neighbors (KNN), linear discriminant analysis (LDA), and random forest classifier (RFC). For balanced accuracy, the black line represents the baseline value, which is the ratio of one class over another.

is increased, meaning that some splits of participants are more generalizable than others.

The accelerometer data provided a meaningful boost in the accuracy of the models,

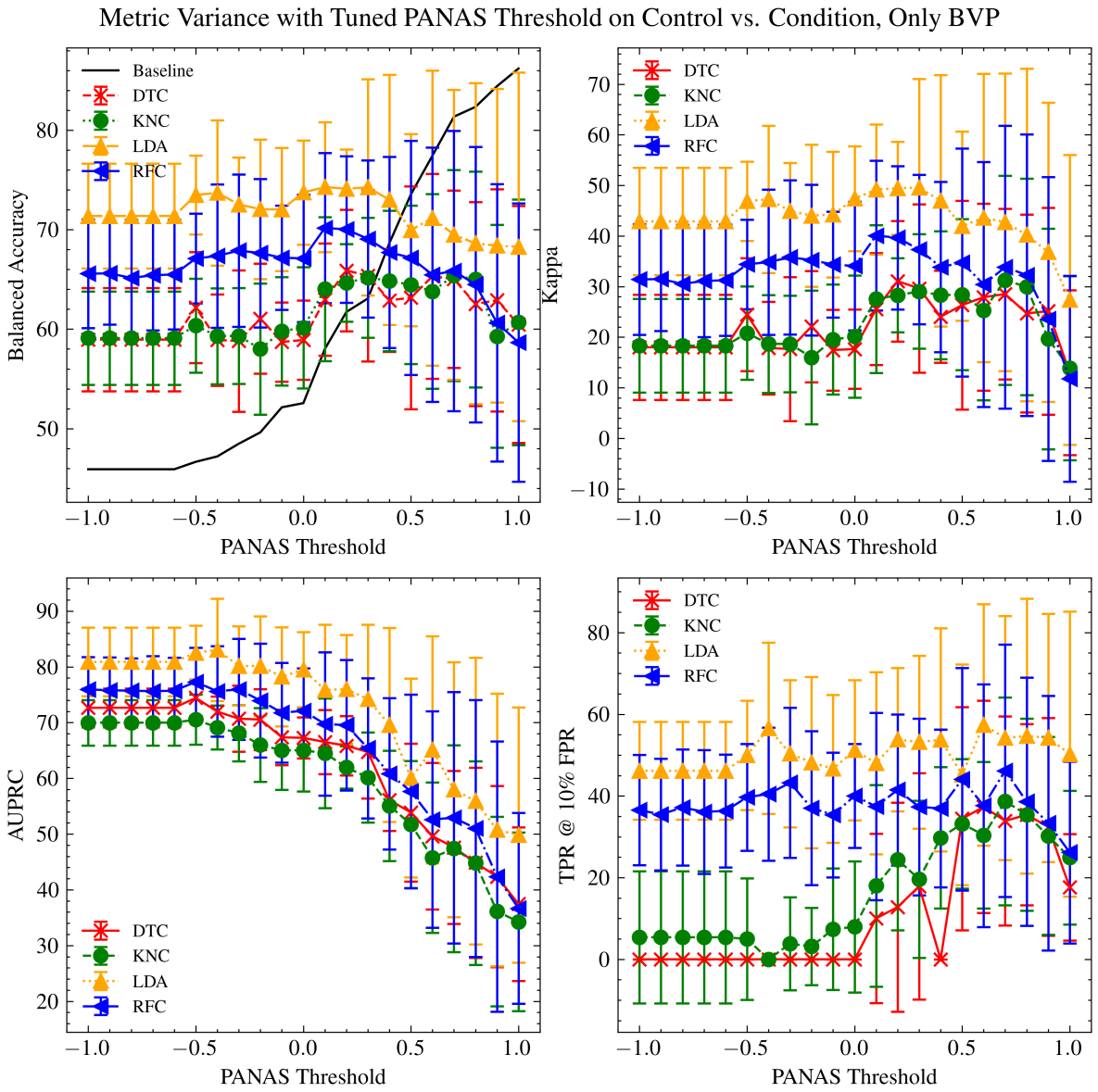

Figure 7. The effect of using various thresholds on the PANAS score to exclude participants from dataset and training to differentiate between intervention conditions using data from the intervention stage, using only BVP. Using the difference between the pre- and post-PANAS positive affect survey values, the PANAS threshold is used to filter out participants that are below the given value. Panels represent different measures of accuracy, including balanced accuracy, Cohen's kappa coefficient, area under the precision recall curve (AUPRC), and true positive rate at 10% of false positive rate (TPR @ 10% FPR). Individual lines represent mean±standard deviation performance for decision tree classifier (DTC), k-nearest neighbors (KNN), linear discriminant analysis (LDA), and random forest classifier (RFC). For balanced accuracy, the black line represents the baseline value, which is the ratio of one class over another.

418 which indicates that petting the dog can be detected in the HAI stress intervention class,
419 which is not present in the control class. This was further reflected in the importance of

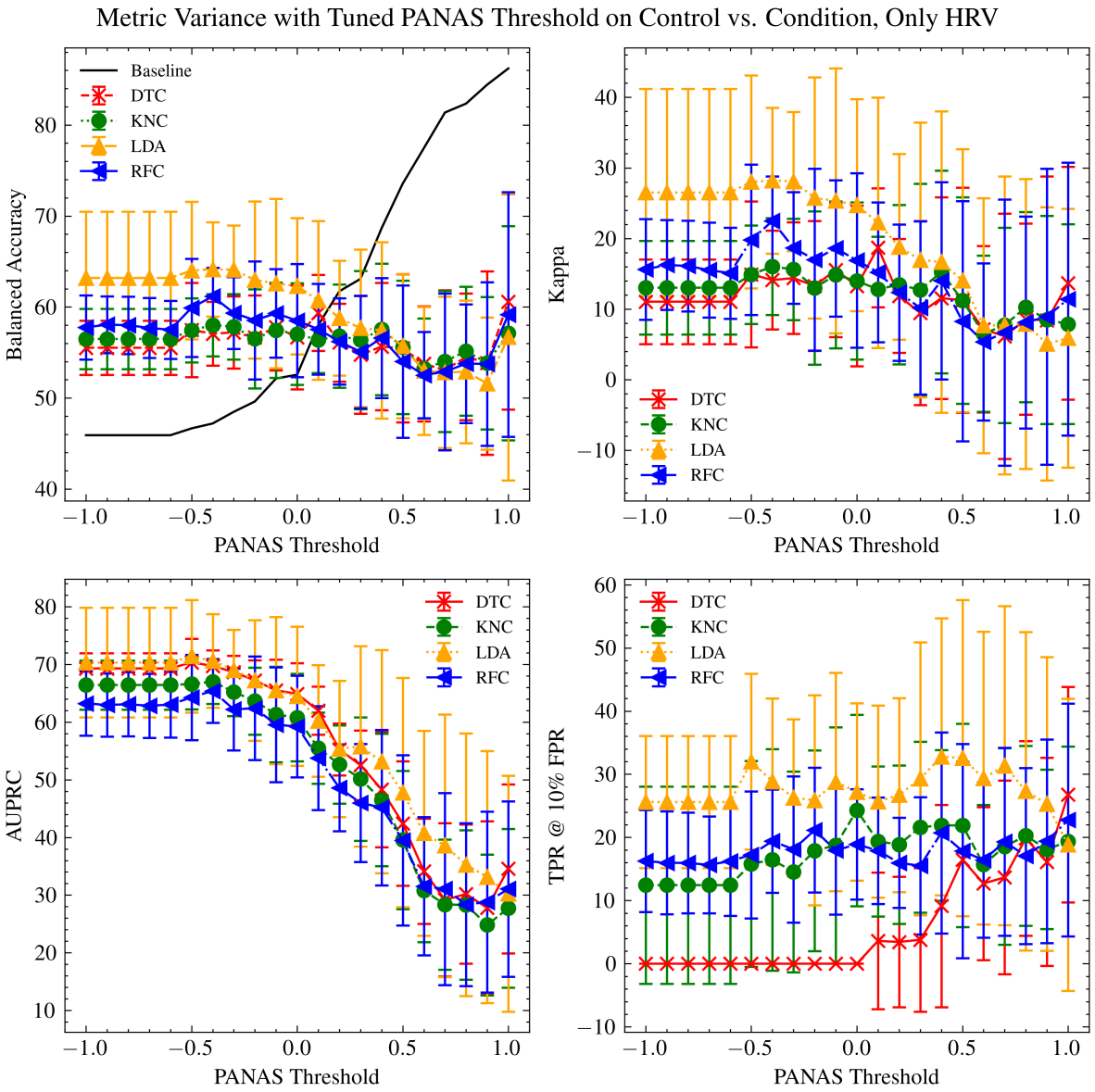

Figure 8. The effect of using various thresholds on the PANAS score to exclude participants from dataset and training to differentiate between intervention conditions using data from the intervention stage, using only HRV. Using the difference between the pre- and post-PANAS positive affect survey values, the PANAS threshold is used to filter out participants that are below the given value. Panels represent different measures of accuracy, including balanced accuracy, Cohen's kappa coefficient, area under the precision recall curve (AUPRC), and true positive rate at 10% of false positive rate (TPR @ 10% FPR). Individual lines represent mean±standard deviation performance for decision tree classifier (DTC), k-nearest neighbors (KNN), linear discriminant analysis (LDA), and random forest classifier (RFC). For balanced accuracy, the black line represents the baseline value, which is the ratio of one class over another.

the permutation feature, which showed that the top models had accelerometry as their top features, and the RFC had all top ten features from accelerometry. Interestingly,

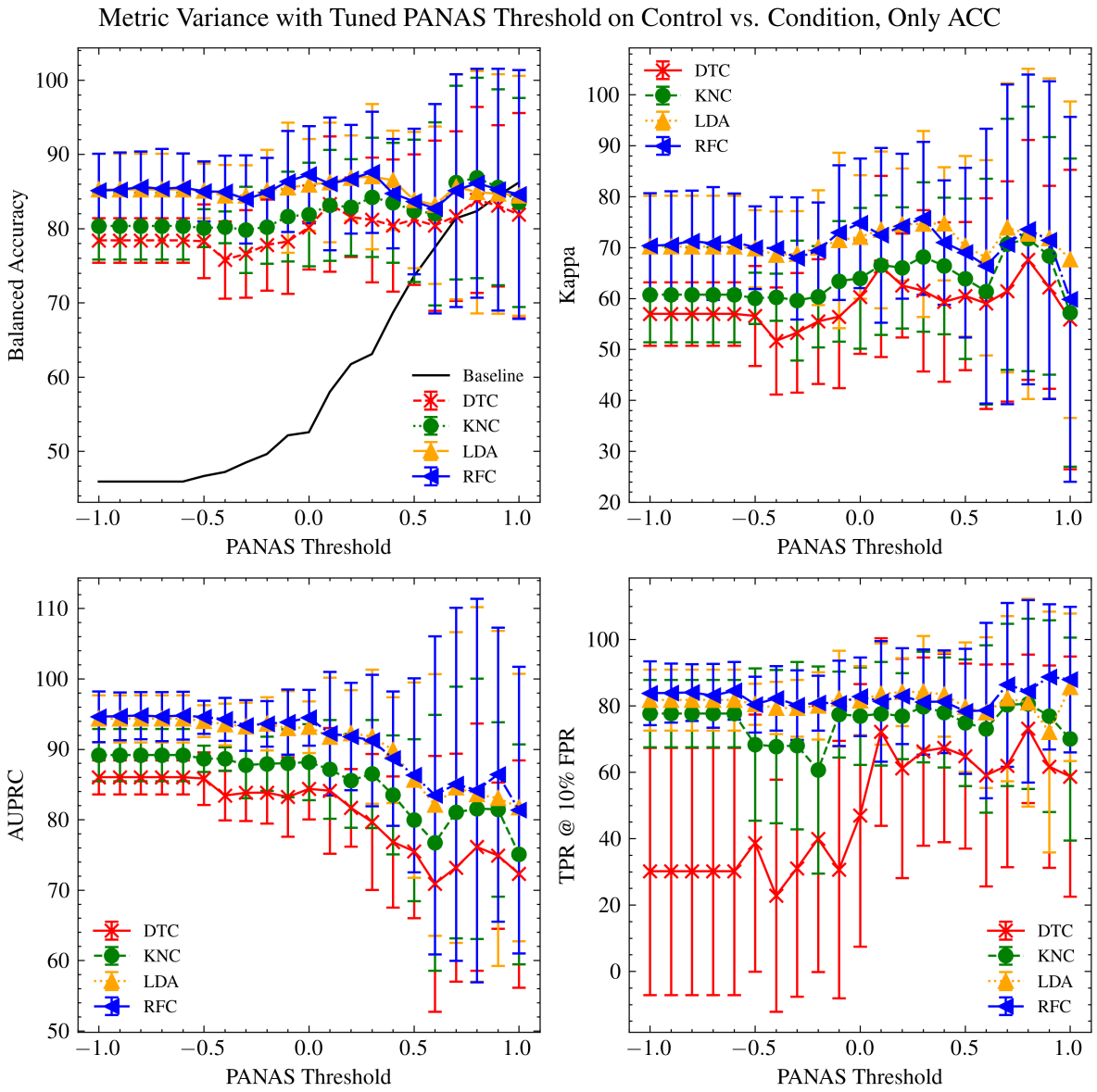

Figure 9. The effect of using various thresholds on the PANAS score to exclude participants from dataset and training to differentiate between intervention conditions using data from the intervention stage, using only ACC. Using the difference between the pre- and post-PANAS positive affect survey values, the PANAS threshold is used to filter out participants that are below the given value. Panels represent different measures of accuracy, including balanced accuracy, Cohen's kappa coefficient, area under the precision recall curve (AUPRC), and true positive rate at 10% of false positive rate (TPR @ 10% FPR). Individual lines represent mean±standard deviation performance for decision tree classifier (DTC), k-nearest neighbors (KNN), linear discriminant analysis (LDA), and random forest classifier (RFC). For balanced accuracy, the black line represents the baseline value, which is the ratio of one class over another.

the RFC model had significant feature correlations, requiring many dozens of features to be randomized before a significant loss in performance. In contrast, LDA could only

sustain the randomization of a handful of features before losing all the prediction power.

When accelerometry was removed, the models had reduced metrics but still outperformed the baseline classification performance. With only EDA, the balanced accuracy was 64.6 ± 5.4 with LDA. With only BVP, the balanced accuracy was 71.4 ± 5.3 with LDA. With only HRV, the balanced accuracy was 63.2 ± 7.3 with LDA. With only accelerometry, the balanced accuracy was 85.3 ± 4.8 with LDA.

Our dataset performs at the same level as other similar datasets. WESAD, with 15 participants, achieves upper 70% and lower 80% top-1 accuracy results across the board, with the DTC model performing the worst [13]. [11] had a wider variety of top-1 accuracy values, but achieved a peak value of 82.3% using their Snake dataset and 68.2% using their cognitive load dataset, each with 23 participants. These two examples shows that CLACIR either outperforms or is on par with other similar datasets. Through the sheer number of participants, our models show generalization in our task that is not able to be explored in other datasets due to limited participant count.

It is possible that other machine learning models could perform better, for instance [11] uses XGBoost to achieve their top performing model on their Snake dataset. Additionally, there is the question of whether models trained with the CLACIR dataset would generalize to other datasets. Does the cognitive load task being learned by the best model also correlate with the social evaluative stress of [13] or the cognitive load of [11]?

The purpose of this study is to make this large physiological dataset available that can be used for affective computing. Moreover, we apply some preliminary machine learning analyses to investigate how the four physiological measures can be used to predict affective state. With CLACIR as a foundation, clinical applications can use machine learning for pilot studies in addition to traditional statistics [2], more robust machine learning models can be incorporated into digital domains like virtual reality [6], or can improve aggression detection approaches [14, 63].

Conclusions

This dataset provides a large sampling of participants with a well-defined protocol that induces a commonly investigated mental state of cognitive load. This contrasts with common physiological datasets, where the main goal is to induce heightened states of emotion through Trier Stress Tests [13] or using provocative video [64]. The other issue is that most datasets are of a small sampling of people, usually less than 40 individuals and more commonly around 15. This leads to methodological issues where the generalization performance of the models cannot be assessed in a fair way. The most common way of assessing generalization performance is the Leave One Subject Out approach, which can introduce bias by selecting the best performing participant to report. By having a very large dataset like CLACIR with 140 individuals, even a 10% train-test split leaves 14 individuals in the test set, which is nearly the same size as WESAD [13].

In addition, this dataset performs well on a variety of machine learning models,

showing that the class labels are aligning well with the physiological data features and that learning is occurring. This is further enhanced by using multiple rich metrics that give insight into other aspects of the model's performance.

Ethical Statement

All data was collected and processed in an ethical manner, in full compliance with the University of Nebraska-Lincoln Internal Review Board's (protocol #19552) and Institutional Animal Care and Use Committee's (protocol #1599) relevant codes of experimentation and legislation. All participants gave written consent to participate and acknowledged that de-identified data could be published publicly.

References

[1] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[2] Laura E Phipps et al. "Assessing physiological arousal and emotional valence during behavioral intervention for pediatric feeding difficulties: A pilot study". In: *Mental Health Science* (2024), e74.

[3] Karla Conn Welch. "Physiological signals of autistic children can be useful". In: *IEEE Instrumentation & Measurement Magazine* 15.1 (2012), pp. 28–32.

[4] Sandra Cano et al. "Wearable solutions using physiological signals for stress monitoring on individuals with autism spectrum disorder (ASD): A systematic literature review". In: *Sensors* 24.24 (2024), p. 8137.

[5] Sarah Sarabadani et al. "Physiological detection of affective states in children with autism spectrum disorder". In: *IEEE Transactions on Affective Computing* 11.4 (2018), pp. 588–600.

[6] Walker Arce and James Gehringer. "Biosensor framework: A C# library for affective computing". In: *Journal of Open Source Software* 6.64 (2021), p. 3455.

[7] Carmen Elisa Orozco-Mora et al. "Stress level estimation based on physiological signals for virtual reality applications". In: *IEEE Access* 10 (2022), pp. 68755–68767.

[8] Nuria Mateos-García et al. "Driver stress detection from physiological signals by virtual reality simulator". In: *Electronics* 12.10 (2023), p. 2179.

[9] Lan-lan Chen et al. "Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers". In: *Expert Systems with Applications* 85 (2017), pp. 279–291.

[10] Surendra Bikram Thapa and Abhijit Sarkar. *Effectiveness of Wearable Devices to Study Driving Stress of Long-haul Truck Drivers in Naturalistic Driving Systems*. Tech. rep. National Surface Transportation Safety Center for Excellence, 2025.

[11] Martin Gjoreski et al. "Datasets for cognitive load inference using wearable sensors and psychological traits". In: *Applied Sciences* 10.11 (2020), p. 3843.

[12] Dimitris Spathis et al. "Learning generalizable physiological representations from large-scale wearable data". In: *arXiv preprint arXiv:2011.04601* (2020).

[13] Philip Schmidt et al. "Introducing wesad, a multimodal dataset for wearable stress and affect detection". In: *Proceedings of the 20th ACM international conference on multimodal interaction*. 2018, pp. 400–408.

[14] Matthew S Goodwin et al. "Predicting aggression to others in youth with autism using a wearable biosensor". In: *Autism research* 12.8 (2019), pp. 1286–1296.

[15] Elise R Thayer and Jeffrey R Stevens. "Effects of human-animal interactions on affect and cognition". In: *Human-Animal Interaction Bulletin* 2022 (2022).

[16] Seyedmajid Hosseini et al. "A multimodal sensor dataset for continuous stress detection of nurses in a hospital". In: *Scientific Data* 9.1 (2022), p. 255.

[17] Ming-Zher Poh, Nicholas C Swenson, and Rosalind W Picard. "A wearable sensor for unobtrusive, long-term assessment of electrodermal activity". In: *IEEE transactions on Biomedical engineering* 57.5 (2010), pp. 1243–1252.

[18] Deniz Ekiz et al. "Can a smartband be used for continuous implicit authentication in real life". In: *IEEE Access* 8 (2020), pp. 59402–59411.

[19] Jeffrey J Walczyk et al. "Advancing lie detection by inducing cognitive load on liars: A review of relevant theories and techniques guided by lessons from polygraph-based approaches". In: *Frontiers in psychology* 4 (2013), p. 14.

[20] GV Portnova et al. "Autonomic and behavioral indicators on increased cognitive loading in healthy volunteers". In: *Neuroscience and Behavioral Physiology* 53.1 (2023), pp. 92–102.

[21] Cornelia Setz et al. "Discriminating stress from cognitive load using a wearable EDA device". In: *IEEE Transactions on information technology in biomedicine* 14.2 (2009), pp. 410–417.

[22] Valentina Markova, Todor Ganchev, and Kalin Kalinkov. "Clas: A database for cognitive load, affect and stress recognition". In: *2019 International Conference on Biomedical Innovations and Applications (BIA)*. IEEE. 2019, pp. 1–4.

[23] Walker Arce and Jeffrey Stevens. *Machine learning classification of stress intervention*. Tech. rep. University of Nebraska-Lincoln, 2022. DOI: 10.13140/RG.2.2.30895.80805.

[24] Henry L Roediger and Kathleen B McDermott. "Creating false memories: Remembering words not presented in lists." In: *Journal of experimental psychology: Learning, Memory, and Cognition* 21.4 (1995), p. 803.

[25] Cathy L McEvoy, Douglas L Nelson, and Takako Komatsu. "What is the connection between true and false memories? The differential roles of interitem associations in recall and recognition." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 25.5 (1999), p. 1177.

[26] J Orbach, Dan Ehrlich, and Helen A Heath. "Reversibility of the Necker cube: I. An examination of the concept of "satiation of orientation"". In: *Perceptual and motor skills* 17.2 (1963), pp. 439–458.

[27] Bernadine Cimprich. "Development of an intervention to restore attention in cancer patients". In: *Cancer nursing* 16.2 (1993), pp. 83–92.

[28] Eva Sahlin et al. "The influence of the environment on directed attention, blood pressure and heart rate—An experimental study using a relaxation intervention". In: *Landscape Research* 41.1 (2016), pp. 7–25.

[29] Marc G Berman, John Jonides, and Stephen Kaplan. "The cognitive benefits of interacting with nature". In: *Psychological science* 19.12 (2008), pp. 1207–1212.

[30] Debra Lynn Rich. "Effects of exposure to nature and plants on cognition and mood: A cognitive psychology perspective". PhD thesis. Cornell University, 2007.

[31] Jonathan D Cohen et al. "Activation of the prefrontal cortex in a nonspatial working memory task with functional MRI". In: *Human brain mapping* 1.4 (1994), pp. 293–304.

[32] Jonathan Peirce et al. "PsychoPy2: Experiments in behavior made easy". In: *Behavior research methods* 51 (2019), pp. 195–203.

[33] David Watson, Lee Anna Clark, and Auke Tellegen. "Development and validation of brief measures of positive and negative affect: the PANAS scales." In: *Journal of personality and social psychology* 54.6 (1988), p. 1063.

[34] Charles D Spielberger et al. *Measuring anxiety and anger with the State-Trait Anxiety Inventory (STAI) and the State-Trait Anger Expression Inventory (STAXI).* Lawrence Erlbaum Associates Publishers, 1999.

[35] David F Cella and Samuel W Perry. "Reliability and concurrent validity of three visual-analogue mood scales". In: *Psychological reports* 59.2 (1986), pp. 827–833.

[36] Mohammed Taj-Eldin et al. "A review of wearable solutions for physiological and emotional monitoring for use by people with autism spectrum disorder and their caregivers". In: *Sensors* 18.12 (2018), p. 4271.

[37] Nishkam Ravi et al. "Activity recognition from accelerometer data". In: *Aaai.* Vol. 5. Pittsburgh, PA. 2005, pp. 1541–1546.

[38] Hoang Minh Thang et al. "Gait identification using accelerometer on mobile phone". In: *2012 International Conference on Control, Automation and Information Sciences (ICCAIS).* IEEE. 2012, pp. 344–348.

[39] Sylvia D Kreibig. "Autonomic nervous system activity in emotion: A review". In: *Biological psychology* 84.3 (2010), pp. 394–421.

[40] Jennifer A Healey and Rosalind W Picard. "Detecting stress during real-world driving tasks using physiological sensors". In: *IEEE Transactions on intelligent transportation systems* 6.2 (2005), pp. 156–166.

[41] Society for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures et al. "Publication recommendations for electrodermal measurements". In: *Psychophysiology* 49.8 (2012), pp. 1017–1034.

[42] Alberto Greco et al. "cvxEDA: A convex optimization approach to electrodermal activity processing". In: *IEEE transactions on biomedical engineering* 63.4 (2015), pp. 797–804.

[43] Mathias Benedek and Christian Kaernbach. "A continuous measure of phasic electrodermal activity". In: *Journal of neuroscience methods* 190.1 (2010), pp. 80–91.

[44] Christian Collet et al. "Autonomic nervous system response patterns specificity to basic emotions". In: *Journal of the autonomic nervous system* 62.1-2 (1997), pp. 45–57.

[45] E Salazar-López et al. "The mental and subjective skin: Emotion, empathy, feelings and thermography". In: *Consciousness and cognition* 34 (2015), pp. 149–162.

[46] Palanisamy Karthikeyan, Murugappan Murugappan, and Sazali Yaacob. "Descriptive analysis of skin temperature variability of sympathetic nervous system activity in stress". In: *Journal of Physical Therapy Science* 24.12 (2012), pp. 1341–1344.

[47] Paul Van Gent et al. "HeartPy: A novel heart rate algorithm for the analysis of noisy signals". In: *Transportation research part F: traffic psychology and behaviour* 66 (2019), pp. 368–378.

[48] Fred Shaffer and Jay P Ginsberg. "An overview of heart rate variability metrics and norms". In: *Frontiers in public health* 5 (2017), p. 258.

[49] Walker Arce. "Unobtrusive Data Collection in Clinical Settings for Advanced Patient Monitoring and Machine Learning". Master of Science thesis in Electrical Engineering. University of Nebraska-Lincoln, 2023.

[50] Simon Föll et al. "FLIRT: A feature generation toolkit for wearable data". In: *Computer Methods and Programs in Biomedicine* 212 (2021), p. 106461.

[51] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[52] Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[53] Debra Wetcher-Hendricks. *Analyzing quantitative data: An introduction for social researchers*. John Wiley & Sons, 2011.

[54] Leo Breiman and Ross Ihaka. *Nonlinear discriminant analysis via scaling and ACE*. Department of Statistics, University of California Davis One Shields Avenue ..., 1984.

[55] L Breiman. "Random forests-random features Technical Report 576". In: *Statistical Department, UC Berkeley, USA* (1999).

[56] Yangguang Liu et al. "A strategy on selecting performance metrics for classifier evaluation". In: *International Journal of Mobile Computing and Multimedia Communications (IJMCMC)* 6.4 (2014), pp. 20–35.

[57] Jesse Davis and Mark Goadrich. "The relationship between Precision-Recall and ROC curves". In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 233–240.

[58] Tom Fawcett. "An introduction to ROC analysis". In: *Pattern recognition letters* 27.8 (2006), pp. 861–874.

[59] S Jothilakshmi and VN Gudivada. "Large scale data enabled evolution of spoken language research and applications". In: *Handbook of Statistics*. Vol. 35. Elsevier, 2016, pp. 301–340.

[60] Sidney D'Mello, Arvid Kappas, and Jonathan Gratch. "The affective computing approach to affect measurement". In: *Emotion Review* 10.2 (2018), pp. 174–183.

[61] Jacob Cohen. "A coefficient of agreement for nominal scales". In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46.

[62] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[63] Walker S Arce et al. "Detecting aggression in clinical treatment videos". In: *Machine Learning with Applications* 14 (2023), p. 100515.

[64] Karan Sharma et al. "A dataset of continuous affect annotations and physiological signals for emotion analysis". In: *Scientific data* 6.1 (2019), p. 196.