

Investigating the Generalizability of Physiological Characteristics of Anxiety

Emily Zhou
Department of Electrical Engineering
University of Southern California
Los Angeles, USA
emilyzho@usc.edu

Mohammad Soleymani
Institute for Creative Technologies
University of Southern California
Los Angeles, USA
msoleyma@usc.edu

Maja J. Matarić
Department of Computer Science
University of Southern California
Los Angeles, USA
mataric@usc.edu

Abstract—Recent works have demonstrated the effectiveness of machine learning (ML) techniques in detecting anxiety and stress using physiological signals, but it is unclear whether ML models are learning physiological features specific to stress. To address this ambiguity, we evaluated the generalizability of physiological features that have been shown to be correlated with anxiety and stress to high-arousal emotions. Specifically, we examine features extracted from electrocardiogram (ECG) and electrodermal (EDA) signals from the following three datasets: Anxiety Phases Dataset (APD), Wearable Stress and Affect Detection (WESAD), and the Continuously Annotated Signals of Emotion (CASE) dataset. We aim to understand whether these features are specific to anxiety or general to other high-arousal emotions through a statistical regression analysis, in addition to a within-corpus, cross-corpus, and leave-one-corpus-out cross-validation across instances of stress and arousal. We used the following classifiers: Support Vector Machines, LightGBM, Random Forest, XGBoost, and an ensemble of the aforementioned models. We found that models trained on an arousal dataset perform relatively well on a previously unseen stress dataset, and vice versa. Our experimental results suggest that the evaluated models may be identifying emotional arousal instead of stress. This work is the first cross-corpus evaluation across stress and arousal from ECG and EDA signals, contributing new findings about the generalizability of stress detection.

I. INTRODUCTION

Anxiety and stress are common responses to actual or perceived threats in daily life. While there are negative connotations to anxiety and stress, they are not necessarily harmful to well-being. Acute anxiety and stress may help us to adapt to stressors and respond accordingly, but severe, prolonged stress may give rise to illness. Chronic stress and anxiety can weaken the immune system and have been linked to many health issues, including major neurological disorders, heart diseases, and high blood pressure [1], [2]. Anxiety disorders and chronic stress are highly prevalent; in the United States alone, with two in three adults report that they have experienced increased stress over the course of the COVID-19 pandemic [3], and rates of moderate to severe anxiety have increased and remain above pre-pandemic levels [4]. Due to the closely related emotional experiences of anxiety and stress [5], we refer to both as “stress” in this paper.

The growing mental health crisis has inspired an increased interest in developing computing tools to aid the early detection of chronic anxiety and stress. However, to the best of

our knowledge, there is less work on generalized methods for anxiety, stress, and affect detection. In this work, we perform cross-corpus stress and emotion recognition using physiological features typically indicative of heightened stress and anxiety across three datasets of stress and general affect in order to examine the generalization properties of these physiological features.

In general, cross-corpus emotion recognition is a more challenging task due to the different types of stimuli, signal quality, and emotion models used in annotation [6], [7], [8]. Studies of the generalization of emotion recognition methods typically classify emotions as high/low arousal or positive/negative valence across different datasets, while generalization of stress detection methods focuses on classifying high/low stress. *Few studies have investigated the generalization of stress detection methods, and to the best of our knowledge, none have attempted to evaluate the generalization of emotion and stress detection methods in a true cross-corpus evaluation*, though prior works have performed leave-one-corpus-out evaluations and/or train-test analyses with data pooled from multiple datasets.

Emotions are often described using the axes of arousal and valence [9], which we describe in greater detail in Section II. Higher stress is indicative of higher arousal, but higher arousal does not necessarily indicate higher stress. In this work, we performed within-corpus, cross-corpus, and leave-one-corpus-out analyses using three affective datasets of stress and arousal to test the following hypothesis:

H1: Stress detection models may be learning features that are general across emotional arousal instead of specific to stress.

The three datasets we used are APD (stress) [10], WESAD (stress) [11], and CASE (arousal) [12]. Because stress is one type of high-arousal emotion, we expect an affect detection model trained on arousal labels to perform relatively well on a previously unseen dataset with stress labels. The converse would verify our hypothesis, *i.e.*, the stress detection models are *not* specific to stress and are instead learning features generalizable to high arousal. This work makes the following contributions:

- 1) The first cross-corpus evaluation across stress and arousal using manually extracted physiological signals

from ECG and EDA signals, contributing new findings about the generalizability of stress detection.

- 2) Regression analysis of the features identified to be indicative of anxiety, stress, and high arousal, comparing their significance and direction of correlation across corpora.

II. RELATED WORK

A. Psychophysiology of Emotional Arousal, Anxiety, and Stress

Among the most commonly used theoretical frameworks for representing emotions is Russell's Circumplex Model of affect [9], which uses the two dimensions of valence (pleasure-displeasure) and arousal (high-low alertness). This model, depicted in Figure 1, proposes that emotions can be represented on perpendicular axes of valence and arousal [13]. Increased emotional arousal is associated with a number of physiological changes caused by autonomic nervous system activity, which include increases in heart rate and the skin conductance response (SCR) amplitude [14] and changes in cutaneous blood flow, piloerection, and sweating, measured via electrocardiogram (ECG) and electrodermal activity (EDA).

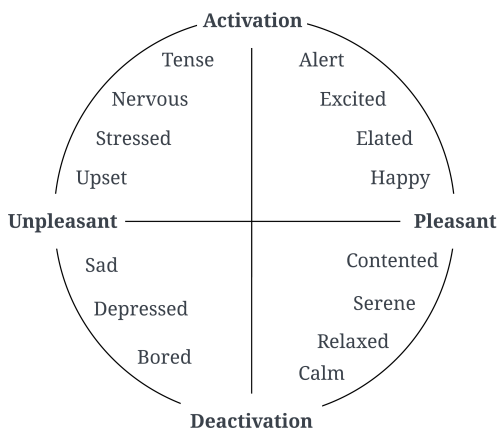


Fig. 1: A graphical representation of the Circumplex Model of affect. The horizontal axis represents the valence dimension and the vertical axis represents the arousal dimension [13].

Under Russell's Circumplex Model, both anxiety and stress are described as high-arousal emotional states. *Anxiety* is defined as a future-oriented mood state of acute sympathetic arousal [15] and is characterized by persistent worries, even in the absence of an external stressor. *Stress* is closely related to anxiety but is typically caused by an external trigger that may be short-term or sustained [16]. Despite this difference, stress and anxiety exhibit the same physiological features that can be derived from ECG and EDA signals, e.g., increased heart rate, decreased heart rate variability, sympathetic dominance measured using frequency metrics of ECG signals, increased skin conductance levels, and increased skin conductance response rates. We describe each of these in greater detail in Section IV-A.

The physiological markers of SNS activity related to anxiety, stress, and arousal can be measured easily and cost-effectively with commercially available sensors, and they have been studied extensively by researchers interested in stress detection. Commonly used physiological signals include electrocardiogram (ECG) and electrodermal activity (EDA) [17]. Heart rate variability and frequency-domain metrics of ECG signals have been used as biomarkers to measure ANS activity, as well as acute and chronic stress levels [17], [18]. Similarly, phasic and tonic (relatively longer-lasting) components of EDA are useful biomarkers for measuring SNS activity and have been used extensively in assessing patients with disorders such as anxiety and depression [19].

B. Anxiety/Stress Detection Using Physiological Signals

Many past works have focused on binary or multi-class classification of stress and/or anxiety and emotion classification using physiological signals from publicly available datasets. They primarily focus on developing better-performing models and identifying the most salient feature sets [20], [21], [22]. Previous work on WESAD [20] used statistical features of ECG, EDA, EMG, and temperature data to perform three-class and binary classification, achieving up to 84.32% and 93.2% accuracy, respectively, with an artificial neural network. Subsequent studies tested those approaches on multiple datasets; Zhu et al. [21] evaluated a stacking ensemble learning model with EDA, ECG, and PPG signals to examine the accuracy of each modality; they achieved the highest accuracy using only EDA features, up to 86.4% with WESAD and 72.3% with the Cognitive Load, Affect, and Stress Database (CLAS). Artificial neural network approaches seeking to achieve better performance than SVM and tree-based methods were also developed [22], [23], [24]. [22] used frequency-domain features as input to a convolutional neural network (CNN) and tested it on the ASCERTAIN, CLAS, MAUS, and WAUC datasets using low, mid, and high-level features of a CNN. They observed accuracy scores of up to 93.58% with ASCERTAIN using a multimodal, multi-level feature fusion approach.

Existing works in stress detection demonstrate the feasibility of using physiological signals from wearable devices to detect stress using SVM, tree-based methods, and artificial neural networks, but they have not determined any ideal combination of modalities or classification techniques. Our approach was guided by previous works that found ML-based classifiers such as SVM, KNN, LightGBM, and XGBoost and an ensemble of those models to be effective due to their high performance and low computational overhead [21]. Previous works found that multimodal feature-level fusion performed better than unimodal features [22]. Following their success, we use multiple modalities, i.e., high-level and statistical features from ECG and EDA signals, and perform a feature-level fusion.

C. Cross-Corpus Affect Recognition

Previous research studying cross-corpus affect recognition primarily focused on general affect recognition using EEG and audiovisual signals. Rayatdoost and Soleymani [25] explored

cross-corpus binary emotion recognition using EEG signals and found that model performance decreased when evaluated across corpora. Zehra et al. [7] performed both within- and cross-corpus binary speech emotion recognition (high vs. low valence) using four datasets with different languages. They found that different classifiers performed better on different datasets, and an ensemble learning approach with majority voting generally improved cross-corpus performance. They achieved an F1-score of 0.63 and accuracy of 63.26% in the cross-corpus setting, training on the Urdu corpus and testing on German. Zhang et al. [8] showed that adding unlabeled speech data to a pooled training set can enhance model performance in both within- and cross-corpus settings. They achieved 66.6% accuracy in binary affect recognition (high vs. low arousal) using a leave-one-corpus-out validation approach with six different datasets.

To the best of our knowledge, Baird et al. [26] are the only authors who have reported a cross-corpus study using physiological signals for stress recognition. Their analysis focused on predicting sequential saliva-based cortisol measures from the FAU-TSST, ULM-TSST, and REG-TSST datasets. However, the biochemical analysis procedure for cortisol extraction varied between datasets, making the derived values not completely comparable. Rather than training a model on one dataset and testing on another to evaluate the generalizability of their approach, they evaluated a joint regression model in which data were pooled from multiple datasets. They found that respiration and heart rate were highly predictive of cortisol levels, and therefore, effective markers of stress. Furthermore, a review of studies on stress detection using physiological signals and wearable sensors between 2012 and 2022 by [17] found no studies attempting to validate a trained model on a completely new, unseen dataset. *To the best of our knowledge, we contribute the first study using physiological signals from wearable devices for cross-corpus and cross-label stress and arousal detection.*

TABLE I: List of abbreviations used

Abbreviation	Refers to
ECG	Electrocardiogram
EDA	Electrodermal activity
HF-RR	Power in the high-frequency band of RR intervals
LF-HF ratio	Ratio of LF to HF power
LF-RR	Power in the low-frequency band of RR intervals
LGBM	LightGBM
RF	Random Forest
RMSSD	Root mean square of successive differences between normal heartbeats
SCL	Skin conductance level
SCR	Skin conductance response
SDNN	Standard deviation of the IBI of normal sinus beats
TSST	Trier Social Stress Test
XGB	XGBoost

III. DATASETS

A. Overview

We performed stress and affect detection using the Anxiety Phases Dataset (APD), Wearable Stress and Affect Detection (WESAD), and the Continuously Annotated Signals of Emotion (CASE) dataset. In our analysis, we focused on ECG and EDA signals due to their availability across all datasets. Each of these datasets contains data from various wearable devices that collect physiological signals and relies on participant self-reports to generate affective labels. Baseline signals were collected during an initial relaxation phase, and recovery signals were collected during the ending "cool-down" phase in each study. Between each phase of emotion elicitation, subjects were given a short rest period before the next stimulus.

The datasets differ primarily in the method of emotion elicitation and type of self-report used. APD targeted bug phobia and public speaking phobia in two tasks and uses the Subjective Units of Distress scale (SUDs) [27] and the Liebowitz Social Anxiety Scale (LSAS) [28]. WESAD used eleven humorous video clips to elicit amusement and the Trier Social Stress Test (TSST) [29], which consists of a public speaking and a mental arithmetic task, to elicit stress. Self-reports were collected using the Positive and Negative Affect Schedule (PANAS) [30], six-item State-Trait Anxiety Inventory (STAI) scored on a four-point Likert scale [31], Self-Assessment Manikins (SAM) [32], and Short Stress State Questionnaire (SSSQ) [33]. CASE introduced a new method of collecting affect self-reports; participants watched video stimuli and reported their continuous emotional experience using a joystick [34], where the x-axis measured valence and the y-axis measured arousal on a scale from 1 to 10. Next, we describe each dataset in greater detail.

B. Anxiety Phases Dataset

APD consists of electrocardiogram (ECG), electrodermal activity (EDA), and accelerometer (ACC) signals from 52 participants in lab-controlled settings across different phases of induced emotions: rest, anticipation, exposure to an anxiety trigger, recovery, and a spoken reflection of the experience. ECG signals were recorded at 250 Hz, and EDA signals were recorded at 50 Hz. Participants wore a Zephyr BioHarness 3.0 on the chest, and a Grove-GSR Sensor on each wrist.

Two anxiety triggers were used: a bug-box task in which participants were asked to release a fake bug from a small box, and a speech anxiety task where participants were asked to rank three topics in order of difficulty, then prepare a 3-minute speech about the most difficult topic. At the end of the study, participants reported their anxiety levels using SUDs and LSAS. Each experiment phase lasted 3 minutes.

C. Wearable Stress and Affect Detection

WESAD is a dataset with stress conditions that includes ECG, EDA, ACC, electromyography (EMG), respiration, and temperature signals collected from 15 subjects across rest, amusement, stress, and two meditation phases. ECG and EDA signals were both recorded at 700 Hz using the chest-worn

RespiBAN Professional and wrist-worn Empatica E4. We used signals collected from the chest-worn sensor because the original paper reported slightly higher accuracies using a single chest-worn sensor as opposed to the wrist-worn sensor. Similarly, [35] found that chest device features had higher importance than wrist-device features in WESAD.

The TSST was used in the stress condition, and a set of 11 humorous video clips was used in the amusement condition. After each condition, subjects self-reported affect using the PANAS, STAI, SAM, and SSSQ questionnaires. The baseline phase was 20 minutes long, the amusement condition was about 6.5 minutes, the stress condition was 10 minutes, and each meditation phase was 7 minutes.

D. Continuously Annotated Signals of Emotion

CASE is a multimodal dataset containing data from 30 participants recorded using ECG, blood volume pulse (BVP), galvanic skin response (GSR, which records the same type of signal as EDA), respiration (RSP), skin temperature (TEMP), and electromyography (EMG) sensors. The authors selected 8 videos previously validated in other studies to elicit amusement, boredom, relaxation, and fear, as well as three additional videos aimed at eliciting a state of relaxation for baseline measurements. Each video lasted 2 to 3 minutes.

IV. METHODOLOGY

A. Data Preprocessing, Segmentation, and Feature Extraction

Our preprocessing methods were modeled after those described in [11] and [10]. ECG and EDA signals were first filtered using preprocessing methods from the biosppy [36] and NeuroKit [37] Python open-source libraries to reduce noise and remove baseline wander. Then, as recommended by [38], for each phase in the datasets, we used 60-second sliding windows with an overlap of 30 seconds to extract the following ECG and EDA metrics from segments of the denoised signal.

Electrocardiogram (ECG):

- **BPM:** Heart rate increases in the face of a stressor [39]. BPM was calculated using the biosppy.ecg module.
- **RMSSD and SDNN:** Decreased HRV is associated with elevated anxiety [40]. RMSSD reflects the beat-to-beat variance in HR and is the primary time-domain measure used to estimate the vagally mediated changes reflected in HRV [41]. Lower SDNN and RMSSD was reported in subjects with various anxiety disorders in [42], [43], and [44]. RMSSD and SDNN were calculated using pyhrv [45].
- **HF-RR and LF-RR:** The high-frequency band of HR oscillations ranges from 0.15–0.4 Hz, and the low-frequency band of HR oscillations ranges from 0.04–0.15 Hz [41]. The HF band reflects parasympathetic influences, and lower HF power is correlated with stress, panic, anxiety, or worry. Increased LF power is associated with anxiety [43]. In general, LF power may be generated by the sympathetic nervous system [41]. The Fast Fourier Transform was used to calculate HF-RR and LF-RR.

- **LF-HF ratio:** Higher LF/HF ratio indicates sympathetic dominance, which occurs during fight-or-flight responses or decreased parasympathetic activity [41]. In healthy individuals, acute stress increases the LF/HF ratio and decreases HF [46]; [40] describes other studies that have reported the same pattern.

Electrodermal Activity (EDA):

- **Mean SCL:** The skin conductance level is the tonic component of EDA. Changes in the SCL are thought to reflect general changes in autonomic arousal [47], and increased SCL is associated with anxiety [19]. SCL was calculated using the NeuroKit Python library, then averaged over the 60-second window.
- **SCR rate:** Skin conductance response is the phasic component of EDA [19] and a higher skin conductance response rate has been associated with anxiety, threat responses, and emotional arousal [48]. SCR rate was also calculated using NeuroKit.

We also extracted the mean, median, standard deviation, and variance from filtered ECG and EDA signals.

B. Label Generation

Each dataset used a different method to collect self-reports, so it was necessary to process those data in a way that allowed for cross-corpus comparison.

APD: Participants' self-reported anxiety levels from the SUDs questionnaire were used to generate binary labels. The version of SUDs used in APD ranges from 0 to 100, and we used 50 as a fixed threshold of the median score, to differentiate high from low anxiety. Participant phases with scores 50 and above were labeled as 1 (high anxiety) and those below 50 were labeled as 0 (low anxiety).

WESAD: Participants' STAI responses were used to generate binary labels. The 6-item short form STAI questionnaire [49] ranges from 6 to 24, and we used the median, 15, as the fixed threshold to determine high vs. low anxiety.

CASE: Participants' self-reports of arousal were used to generate binary labels. Since CASE consists of continuous self-reports, we averaged the self-reported arousal values for each phase to be consistent with APD and CASE. The median value of the joystick range, 5, was used as the fixed threshold.

C. Statistical Analysis and Model Training

We first used a linear mixed effects model from the statsmodels Python library [50] to examine the effects of the selected physiological features on stress and emotional arousal within each dataset. We set the physiological features as fixed effects and the subject group as random effects.

Next, we conducted experiments in within-corpus, cross-corpus, and leave-one-corpus-out settings. Our results from the within-corpus condition are used as baselines for the cross-corpus and LOCO experiments. Five machine learning techniques were used to perform binary classification: Support Vector Machine (SVM), LightGBM (LGBM) [51], Random Forest (RF), XGBoost (XGB), and an ensemble of the previous four models. These models were chosen based on the

SOTA performance achieved by SVM, KNN, and tree-based classifiers in previous affect detection studies, discussed in Section II. We evaluated both majority voting and equally weighted average ensembles of models. The equally weighted ensemble approach performed significantly better, so we omit the majority voting results.

In each experimental setting, data samples were generated from the extracted features by segmenting each phase into 2-minute chunks and calculating the mean for each chunk. Rather than take the mean value over the whole phase, we chose 2-minute segments, since the shortest phase across all three datasets was 2 minutes long. This process was repeated for each feature and phase for each participant. Then, the features for each phase were concatenated to form the input feature vector, which was then normalized before model training. In cases of moderate class imbalance, SMOTE [52] was used. Model hyperparameters were selected using a simple grid search algorithm.

In the cross-corpus and leave-one-corpus-out (LOCO) experiments, we used the following definitions of binary classification: when testing on APD or WESAD, it is defined as classifying an individual as having either high or low anxiety, and when testing on CASE, it is defined as classifying an individual as having either high or low emotional arousal, regardless of the labels in the training dataset used.

Within-corpus: In the within-corpus case, models were first trained and tested within datasets, using a 5-fold cross-validation technique. The participants were split into 5 equally sized groups, and data instances from each of the phases were aggregated to form each fold.

Cross-corpus: In the cross-corpus setting, we trained and tested classifiers across pairs of datasets. One dataset was used as the training set, while the other was held out as the test set to evaluate the generalizability of the features and models.

Leave-one-corpus-out: Finally, we conducted a LOCO experiment using SVM, LGBM, RF, XGB, and an ensemble of the four classifiers. Two datasets were pooled and used as input to the model, while the remaining dataset was held out as the test set.

V. RESULTS AND DISCUSSION

A. Regression Analysis

Results from the regression analysis are reported in Table II. Heart rate (BPM) was significant across all three datasets, while SCR rate was significant and positively correlated in both APD and WESAD. We found that half of the features had the same direction of correlation across all three datasets. BPM, HF-RR, LF-RR, the median of ECG, and the median of EDA were positively correlated to a positive label, and SDNN, the mean and variance of ECG, and the mean of EDA were negatively correlated. Several of these correlations are consistent with previous work in psychology and affect detection: increased heart rate [39] and power in the low-frequency band [43] have been identified as high-level features indicative of SNS activity. SDNN is a measure of HRV, and decreased HRV is known to be correlated with increased

stress [40]. While SCR rate has been found to be positively correlated with stress and anxiety [48], we find a slightly positive correlation in WESAD and CASE. This could be due to the elicitation of amusement in both datasets, which was found to have a lower skin conductance than fear in [53]. *Given that the majority of physiological features exhibited the same direction of correlation in all three datasets, the statistical analysis results support the generalizability of these features across stress and anxiety.*

B. Within-Corpus Performance

We use accuracy and AUC metrics to analyze model performance in the within-corpus, cross-corpus, and LOCO settings. Table III presents our results in the within-dataset condition, as well as the results achieved with a random classifier to serve as a baseline for the within-dataset experiment. All models performed significantly better than a random classifier, and we achieve SOTA performance on WESAD in the within-dataset setting compared to previous binary classification studies [20], [21] using an ensemble of models. The ensemble of models performed the best overall across all datasets. It is important to note that our goal in the within-dataset case was not to achieve SOTA performance across all datasets. This step verifies that our pipeline, from preprocessing to model training and classification, is a viable method of affect detection for the APD, WESAD, and CASE datasets. Model performance is the lowest in APD, which may be attributable to more noise present in the data. These results verify that our preprocessing, feature extraction, and classification methods are viable for detecting stress and arousal.

C. Specificity of Stress Detection: Cross-Corpus and Leave-One-Corpus-Out Performance

Cross-corpus results are presented in Table IV, and LOCO results in Table V. We found that model performance in cross-corpus stress and affect detection was lower than within-corpus performance, consistent with the findings in [25]. However, we still achieved better-than-chance accuracy and AUC scores across all pairs of datasets.

Our goal was to identify whether stress detection models and the selected physiological features are specific to stress or generalizable to arousal using cross-corpus analyses across three datasets. The cross-corpus and LOCO experiments were designed to evaluate the specificity of stress detection by testing models with unseen data and different labels (*i.e.*, stress vs arousal) from a completely different distribution. In the case where both training and testing datasets were labeled with stress (APD/WESAD), we achieved up to 68.3% accuracy and an AUC score of 0.761 testing on WESAD, and 56.9% accuracy and an AUC score of 0.561 testing on APD. Some train-test instances were not much better than random guessing, *e.g.* testing on APD regardless of the training dataset in both cross-corpus and LOCO cases. This is unsurprising given the lower performance for APD in the within-corpus case; the discrepancy may be due to a higher signal-to-noises ratio present in APD data.

TABLE II: Correlation coefficients, p -values, and R -squared values (conditional/marginal) obtained from a mixed linear model fitted to each dataset. An asterisk (*) indicates significance, i.e., a two-tailed p -value of less than 0.1, or equivalently, a one-tailed p -value of less than 0.05.

	APD		WESAD		CASE	
	R-squared: 0.688/0.120		R-squared: 0.797/0.556		R-squared: 0.629/0.325	
	Coef	p-value	Coef	p-value	Coef	p-value
BPM	0.239	2.32E-12*	0.441	1.01E-79*	0.254	2.47E-07*
RMSSD	-0.013	0.771	0.17	1.54E-13*	0.087	0.027*
HF_RR	1.048	0.360	0.053	0.486	0.645	0.007*
LF_RR	-1.286	0.268	0.033	0.689	-0.769	0.002*
SDNN	-0.015	0.713	-0.105	3.18E-06*	-0.032	0.339
Mean SCL	1.74	0.338	-0.037	0.043*	0.349	9.90E-13*
SCR rate	0.204	0.1*	-0.066	3.55E-06*	-0.034	0.304
LF/HF ratio	0.055	0.153	0.121	3.20E-06*	0.049	9.64E-08*
ECG mean	-1.275	0.037*	-0.148	0.558	-1.694	0.144
ECG median	1.295	0.037*	0.134	0.601	1.653	0.154
ECG std	-0.075	0.306	0.038	0.688	1.024	2.56E-24*
ECG var	-0.041	0.559	-0.157	0.07*	-0.936	3.44E-21*
EDA mean	-1.82	0.318	-0.03	0.007*	-0.594	0.001*
EDA median	0.019	0.894	0.063	0.007*	0.234	0.162
EDA std	0.058	0.246	0.065	0.06*	-0.094	0.107
EDA var	-0.041	0.406	-0.024	0.419	0.134	0.017*

TABLE III: Within-corpus classification results for stress vs. non-stress (Acc and AUC score) in APD and WESAD and high vs low arousal in CASE. We use SVM, LightGBM, Random Forest, XGBoost, and an ensemble of the previous models. The best results across all classifiers are bolded.

Model	Acc	AUC	Acc	AUC	Acc	AUC
	APD		WESAD		CASE	
Random	0.471	0.467	0.509	0.432	0.471	0.503
SVM	0.549	0.498	0.860	0.720	0.800	0.658
LGBM	0.547	0.529	0.836	0.720	0.788	0.661
RF	0.553	0.540	0.845	0.735	0.792	0.692
XGB	0.561	0.528	0.842	0.775	0.767	0.671
Ensemble	0.645	0.608	0.990	0.969	0.813	0.736

TABLE IV: Cross-corpus classification results using SVM, LightGBM, Random Forest, XGBoost, and an ensemble of the previous models (accuracy and AUC score). Random classifier performance is included as a point of comparison.

Model	Acc	AUC	Acc	AUC	Acc	AUC
	APD/WESAD		APD/CASE		WESAD/CASE	
Random	0.509	0.432	0.471	0.503	0.471	0.503
SVM	0.358	0.530	0.458	0.500	0.463	0.510
LGBM	0.434	0.597	0.610	0.611	0.696	0.617
RF	0.260	0.515	0.515	0.564	0.598	0.582
XGB	0.499	0.641	0.569	0.644	0.610	0.626
Ensemble	0.658	0.709	0.708	0.611	0.644	0.600
	WESAD/APD		CASE/APD		CASE/WESAD	
Random	0.471	0.467	0.471	0.503	0.509	0.432
SVM	0.552	0.539	0.535	0.522	0.774	0.649
LGBM	0.550	0.542	0.546	0.537	0.680	0.555
RF	0.513	0.527	0.538	0.544	0.559	0.637
XGB	0.503	0.550	0.536	0.529	0.755	0.605
Ensemble	0.569	0.561	0.567	0.551	0.759	0.591

To evaluate our hypothesis, we examine the cross-label training results, i.e., training on a stress dataset and testing on an arousal dataset, and vice versa. Since stress is an example of a high arousal emotion, we expected models trained on arousal labels to perform relatively well on new instances with stress labels, and models trained on stress labels to be specific

TABLE V: Leave-one-corpus out classification results using SVM, LightGBM, Random Forest, XGBoost, and an ensemble of the previous models (accuracy and AUC score).

Model	Acc	AUC	Acc	AUC	Acc	AUC
	Test: APD		Test: WESAD		Test: CASE	
Random	0.471	0.467	0.509	0.432	0.471	0.503
SVM	0.501	0.492	0.707	0.510	0.621	0.546
LGBM	0.505	0.497	0.590	0.475	0.535	0.534
RF	0.515	0.513	0.753	0.536	0.463	0.562
XGB	0.516	0.507	0.182	0.571	0.492	0.541
Ensemble	0.521	0.506	0.712	0.506	0.610	0.552

to stress and therefore perform worse on a dataset labeled with arousal. Instead, we found that training on APD and testing on CASE resulted in better performance (accuracies up to 70.8% and AUC scores up to 0.644) than training on CASE and testing on APD (accuracies up to 56.7% and AUC scores up to 0.551). Training on WESAD and testing on CASE resulted in accuracies up to 69.6% and AUC scores up to 0.617. CASE/APD achieved a maximum accuracy of 77.4% and AUC score of 0.649.

Therefore, our results are supportive of H1: *emotion classification models trained on stress and tested on arousal labels can perform nearly as well or better than training on arousal and testing on stress. In fact, cross-label evaluations resulted in higher accuracies and AUC scores than training and testing between stress datasets.* It appears that the stress detection models we tested may be learning features *general across emotional arousal* instead of *specific to stress*. While our early work in the specificity of stress features is limited in that we did not perform explainable AI techniques such as the calculation of Shapley values, our results suggest that this is an important direction for future work in stress detection.

VI. CONCLUSION

Our work contributes the first study of cross-corpus stress and affect detection across datasets labeled with stress and

emotional arousal. We extracted physiological features from ECG and EDA signals and performed stress and arousal detection in within-corpus, cross-corpus, and leave-one-corpus-out experiments on three affective datasets. Our results show that stress recognition models are able to recognize emotional arousal better than stress in difficult cross-corpus settings with different emotional stimuli, suggesting that such recognition models may be detecting an emotionally aroused state that is more general than stress. A limitation of this work is the relatively small number of datasets and features used. We aimed to use common high-level features of ECG and EDA signals that have been found to be indicative of ANS activity, but deep learning-based features are also highly useful in detecting stress and anxiety, which can be explored in future cross-corpus work. We also did not identify a set of physiological features that are specific to stress/anxiety, which merits further investigation with metrics and tools such as Shapley values and feature importances. However, our results suggest the importance of cross-corpus and cross-label experiments to verify that features learned by stress detection model are specific to stress.

This work motivates a deeper evaluation of stress detection and the development of models that use features specific to stress to accurately identify only stressful instances. To support the improvement of stress detection models toward in-the-wild deployments, more work is needed in cross-corpus settings using larger datasets and more high-arousal emotions. Future work may benefit from collecting physiological signals from a wider variety of stressors and in-the-wild environments to further explore the generalizability of stress detection across settings.

ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation under Grant No. 2211550 and the University of California's Center for Undergraduate Research in Viterbi Engineering (CURVE) fellowship. The authors thank Prof. Bhaskar Krishnamachari for his insights and expertise throughout the development and exploratory analyses of this work.

REFERENCES

- [1] N. Schneiderman, G. Ironson, and S. D. Siegel, "Stress and health: Psychological, behavioral, and biological determinants," *Annual Review of Clinical Psychology*, vol. 1, no. 1, p. 607–628, 2005.
- [2] A. Mariotti, "The effects of chronic stress on health: New insights into the molecular mechanisms of brain–body communication," *Future Science OA*, vol. 1, no. 3, 2015.
- [3] "What is Stress - Stress in America 2022," Mar 2022. [Online]. Available: <https://www.stress.org/daily-life>
- [4] "Mental Health and COVID-19 2021 Data," 2021. [Online]. Available: <https://mhanational.org/mental-health-and-covid-19-april-2022-data>
- [5] N. Daviu, M. R. Bruchas, B. Moghaddam, C. Sandi, and A. Beyeler, "Neurobiological links between stress and anxiety," *Neurobiology of Stress*, vol. 11, 2019.
- [6] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Using multiple databases for training in emotion recognition: To unite or to vote?" 08 2011, pp. 1553–1556.
- [7] W. Zehra, A. R. Javed, Z. Jalil, H. U. Khan, and T. R. Gadekallu, "Cross corpus multi-lingual speech emotion recognition using ensemble learning," *Complex & Intelligent Systems*, vol. 7, no. 4, p. 1845–1854, 2021.
- [8] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, 2011, pp. 523–528.
- [9] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.
- [10] H. Senaratne, L. Kuhlmann, K. Ellis, G. Melvin, and S. Oviatt, "A multimodal dataset and evaluation for feature estimators of temporal phases of anxiety." New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: <https://doi.org/10.1145/3462244.3479900>
- [11] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection." New York, NY, USA: Association for Computing Machinery, 2018. [Online]. Available: <https://doi.org/10.1145/3242969.3242985>
- [12] K. Sharma, C. Castellini, E. L. van den Broek, A. Albu-Schaeffer, and F. Schwenker, "A dataset of continuous affect annotations and physiological signals for emotion analysis," *Scientific Data*, 2019.
- [13] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and Psychopathology*, 2005.
- [14] F. D'Hondt, M. Lassonde, O. Collignon, A.-S. Dubarry, M. Robert, S. Rigoulot, J. Honore, F. Lepore, and H. Sequeira, "Early brain-body impact of emotional arousal," *Frontiers in Human Neuroscience*, 2010.
- [15] J. M. Cisler, B. O. Olatunji, M. T. Feldner, and J. P. Forsyth, "Emotion regulation and the anxiety disorders: An integrative review," Mar 2010. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2901125/>
- [16] M. Alvord and R. Halfond, "What's the difference between stress and anxiety?" Oct 2019. [Online]. Available: <https://www.apa.org/topics/stress/anxiety-difference>
- [17] G. Vos, K. Trinh, Z. Sarnyai, and M. R. Azghadi, "Generalizable machine learning for stress monitoring from wearable devices: A systematic literature review," Feb 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1386505623000436>
- [18] A. Sepúlveda, F. Castillo, C. Palma, and M. Rodríguez-Fernández, "Emotion recognition from ECG signals using wavelet scattering and machine learning," May 2021. [Online]. Available: <https://www.mdpi.com/2076-3417/11/11/4945>
- [19] N. Ali, H. Tschenett, and U. M. Nater, "Biomarkers of stress and disease," Dec 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/abs/pii/B9780323914970002319>
- [20] P. Bobade and M. Vani, "Stress Detection with Machine Learning and Deep Learning using Multimodal Physiological Data," in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2020, pp. 51–57.
- [21] L. Zhu, P. Spachos, and S. Gregori, "Multimodal Physiological Signals and Machine Learning for Stress Detection by Wearable Devices," in *2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2022, pp. 1–6.
- [22] R. Kuttala, R. Subramanian, and V. R. M. Oruganti, "Multimodal Hierarchical CNN Feature Fusion for Stress Detection," *IEEE Access*, vol. 11, pp. 6867–6878, 2023.
- [23] C.-P. Bara, M. Papakostas, and R. Mihalcea, "A deep learning approach towards multimodal stress detection," in *AffCon AAAI*, 2020, pp. 67–81.
- [24] F. Albertetti, A. Simalastar, and A. Rizzotti-Kaddouri, "Stress detection with deep learning approaches using physiological signals," in *IoT Technologies for HealthCare*. Springer International Publishing, 2021.
- [25] S. Rayatdoost and M. Soleymani, "Cross-corpus EEG-based Emotion Recognition," in *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2018, pp. 1–6.
- [26] A. Baird, A. Triantafyllopoulos, S. Zänkert, S. Ottl, L. Christ, L. Stappen, J. Konzok, S. Sturmbauer, E.-M. Meßner, B. M. Kudielka, N. Rohleder, H. Baumeister, and B. W. Schuller, "An Evaluation of Speech-Based Recognition of Emotional and Physiological Markers of Stress," *Frontiers in Computer Science*, vol. 3, 2021. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fcomp.2021.750284>
- [27] "Subjective units of distress scale." p. 361, 2015.

- [28] R. G. Heimberg, K. J. Horner, H. R. Juster, S. A. Safren, E. J. Brown, F. R. Schneir, and M. R. Liebowitz, "Psychometric properties of the liebowitz social anxiety scale," *Psychological Medicine*, vol. 29, no. 1, p. 199–212, 1999.
- [29] C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer, "The 'trier social stress test' – a tool for investigating psychobiological stress responses in a laboratory setting," *Neuropsychobiology*, vol. 28, no. 1–2, p. 76–81, 1993.
- [30] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: The PANAS scales," *Journal of Personality and Social Psychology*, vol. 54, no. 6, p. 1063–1070, 1988.
- [31] B. M. Barker, H. R. Barker Jr., and A. P. Wadsworth Jr., "Factor analysis of the items of the state-trait anxiety inventory," *Journal of Clinical Psychology*, vol. 33, no. 2, pp. 450–455, 1977.
- [32] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/0005791694900639>
- [33] W. S. Helton, "Validation of a short stress state questionnaire," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 48, no. 11, pp. 1238–1242, 2004. [Online]. Available: <https://doi.org/10.1177/154193120404801107>
- [34] K. Sharma, C. Castellini, F. Stulp, and E. L. van den Broek, "Continuous, real-time emotion annotation: A novel joystick-based analysis framework," *IEEE Transactions on Affective Computing*, vol. 11, no. 1, pp. 78–84, 2020.
- [35] J. Lin, S. Pan, C. S. Lee, and S. Oviatt, "An Explainable Deep Fusion Network for Affect Recognition Using Physiological Signals," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. Association for Computing Machinery, 2019, p. 2069–2072. [Online]. Available: <https://doi.org/10.1145/3357384.3358160>
- [36] C. Carreiras, A. P. Alves, A. Lourenço, F. Canento, H. Silva, A. Fred et al., "BioSPPy: Biosignal processing in Python," 2015–, [Online; accessed 2024]. [Online]. Available: <https://github.com/PIA-Group/BioSPPy/>
- [37] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen, "NeuroKit2: A python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, no. 4, pp. 168–1696, feb 2021. [Online]. Available: <https://doi.org/10.3758%2Fs13428-020-01516-y>
- [38] R. Fang, R. Zhang, E. Hosseini, A. M. Parenteau, S. Hang, S. Rafatirad, C. E. Hostinar, M. Orooji, and H. Homayoun, "Prevent over-fitting and redundancy in physiological signal analyses for stress detection," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2022, pp. 2585–2588.
- [39] J. Noteboom, K. Barnholt, and R. Enoka, "Activation of the arousal response and impairment of performance increase with anxiety and stressor intensity," 2001. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/11641349/>
- [40] S. D. Kreibitz, "Autonomic nervous system activity in emotion: A review," Apr 2010. [Online]. Available: <https://www.sciencedirect.com/science/article/abs/pii/S0301051110000827>
- [41] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," Sep 2017. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fpubh.2017.00258/full>
- [42] C. Licht, E. de Geus, R. van Dyck, and P. BW, "Association between anxiety disorders and heart rate variability in the netherlands study of depression and anxiety (nesda)," 2009. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/19414616/>
- [43] J. Prasko, K. Latalova, T. Diveky, A. Grambal, J. Kamaradova, H. Velartova, J. Salinger, J. Opavsky, and P. Silhan, Nov 2011. [Online]. Available: <https://www.nel.edu/userfiles/articlesnew/NEL320511A16.pdf>
- [44] J. A. Chalmers, D. S. Quintana, M. J.-A. Abbott, and A. H. Kemp, "Anxiety disorders are associated with reduced heart rate variability: A meta-analysis," Jul 2014. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4092363/>
- [45] P. Gomes, P. Margaritoff, and H. Silva, "pyHRV: Development and evaluation of an open-source python toolbox for heart rate variability (HRV)," in *Proc. Int'l Conf. on Electrical, Electronic and Computing Engineering (IcETRAN)*, 2019, pp. 822–828.
- [46] M. Pagani, N. Montano, A. Porta, A. Malliani, F. M. Abboud, C. Birkett, and V. K. Somers, "Relationship between spectral components of cardiovascular variabilities and direct measures of muscle sympathetic nerve activity in humans," *Circulation*, vol. 95, no. 6, pp. 1441–1448, 1997. [Online]. Available: <https://www.ahajournals.org/doi/abs/10.1161/01.CIR.95.6.1441>
- [47] J. J. Braithwaite, D. G. Watson, R. Jones, and M. Rowe, 2015. [Online]. Available: <https://www.birmingham.ac.uk/Documents/collegeles/psych/saal/guide-electrodermal-activity.pdf>
- [48] M. G. Craske, A. M. Waters, E. M. Ornitz, H. Negoro, O. V. Lipp, B. Naliboff, and L. R. Bergman, "is aversive learning a marker of risk for anxiety disorders in children?," pp. 954–967, 2008. [Online]. Available: <https://doi.org/10.1016/j.brat.2008.04.011>
- [49] T. M. Marteau and H. Bekker, "The development of a six-item short-form of the state scale of the spielberger state–trait anxiety inventory (stai)," *The British journal of clinical psychology*, vol. 31, no. 3, pp. 301–306, 1992. [Online]. Available: <https://doi.org/10.1111/j.2044-8260.1992.tb00997.x>
- [50] S. Seabold and J. Perktold, "statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.
- [51] H. Zhang, S. Si, and C.-J. Hsieh, "Gpu-acceleration for large-scale tree boosting," 2017.
- [52] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- [53] H. Leng, Y. Lin, and L. A. Zanzi, "An experimental study on physiological parameters toward driver emotion recognition," in *Ergonomics and Health Aspects of Work with Computers*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 237–246.