Comparing the Predictability of Sensor Modalities to Detect Stress from Wearable Sensor Data

Ryan Holder
Washington State University
Pullman, WA, USA
ryan.holder@wsu.edu

Ramesh Kumar Sah
Washington State University
Pullman, WA, USA
ramesh.sah@wsu.edu

Michael Cleveland
Washington State University
Pullman, WA, USA
michael.cleveland@wsu.edu

Hassan Ghasemzadeh
Washington State University
Pullman, WA, USA
hassan.ghasemzadeh@wsu.edu

***Abstract*—Detecting stress from wearable sensor data enables those struggling with unhealthy stress coping mechanisms to better manage their stress. Previous studies have investigated how mechanisms for detecting stress from sensor data can be optimized, comparing alternative algorithms and approaches to find the best possible outcome. One strategy to make these mechanisms more accessible is to reduce the number of sensors that wearable devices must support. Reducing the number of sensors will enable wearable devices to be a smaller size, require less battery, and last longer, making use of these wearable devices more accessible. To progress towards this more convenient stress detection mechanism, we investigate how learning algorithms perform on singular modalities and compare the outcome with results from multiple modalities. We found that singular modalities performed comparably or better than combined modalities on two stress-detection datasets, suggesting that there is promise for detecting stress with fewer sensor requirements. From the four modalities we tested, acceleration, blood volume pulse, and electrodermal activity, we saw acceleration and electrodermal activity to stand out in a few cases, but all modalities showed potential. Our results are acquired from testing with random holdout and leave-one-subject-out validation, using several machine learning techniques. Our results can inspire work on optimizing stress detection with singular modalities to make the benefits of these detection mechanisms more convenient.**

Keywords—stress detection, machine learning, wearable sensors

I. Introduction

In many individuals, stress can lead to unhealthy behaviors as they attempt to relieve that stress. Many individuals resort to harmful substances like alcohol to alleviate their stress. By finding better ways to recognize stress, we hope to improve the mechanisms by which we can lead individuals to healthier responses to their stress.

While it is feasible for a stress recognition system to yield high performance when multiple sensor modalities are available, [1], [2], this approach is often computationally expensive. Prioritizing and reducing the sensor modalities we use can reduce the necessary computational power of our recognition devices. Reducing the computational requirement can enable smaller devices that require less power to perform stress recognition. These smaller devices can be made less costly and can operate longer on a single charge, making them more appealing to the consumer. By investigating the performance of fewer modalities, we can find methods to effectively recognize stress with lighter data storage and computational power requirements. We anticipate that the results we find will lead to stress-recognizing devices that are cheaper and more accessible to those who could benefit from them.

This work was supported in part by the National Science Foundation under numbers IIS-1852163, CNS-1750679, and the Alcohol and Drug Research Program of Washington State University through grant funding to MJC. This investigation was supported in part by funds provided for medical and biological research by the State of Washington Initiative Measure No. 171. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations.

II. Related Work

Stress is a problem that exists in the lives of all people. As a result, researchers have investigated how well we can detect that stress in order to prevent it or advise individuals on how to deal with it. In one such study, Kyriakou et al. [3] used a rule-based mechanism with weights and critical values, based on electrodermal activity (galvanic skin response) and skin temperature, to detect stress. In this study, an aural stimulus was used to stimulate a stress response. These researchers found promising results toward detecting stress with rules that monitor electrodermal activity or skin temperature changes. In a similar study, Sagbas et al. [1] performed stress detection analysis based on smartphone keyboard typing behavior. They created an app that collected accelerometer, gyroscope, and touchscreen interaction data while users typed on their phone screens. This group then compared the results of supervised learning algorithms, including a decision tree, a Bayesian network, and a nearest-neighbor algorithm, to investigate how stress can be detected with smartphone sensors. While these studies examined different datasets, both found promising results detecting stress.

In our paper, we will perform analysis using the Wearable Stress and Affect Detection (WESAD) and the Alcohol and Drug Abuse Research Program (ADARP) datasets. The WESAD dataset has become very common within the realm of stress detection, as it contains public wearable sensor data in a simulated stress environment. Several previous works performed similar machine learning-based analyses on the WESAD dataset, with a goal of finding an optimal means of detecting stress. The purpose of the study by Bobade et al. [2] was to algorithmically identify the best-performing machine learning algorithm for these data among seven choices: random forest, decision tree, AdaBoost, kNN, linear discriminant analysis, SVM, and a deep network. They observed that the multi-layer neural network performed best on these data. In many of these stress detection investigations, researchers have

focused on electrodermal activity (EDA), or galvanic skin response (GSR), as an indicator or stress. Aqajari et al. focused on automatically and manually extracting features from the electrodermal activity data that yielded optimal predictive performance [4]. A similar study from Hsieh et al. [5] focused on how novel features could be extracted from electrodermal activity to optimize machine learning results.

Alongside the stress detection issue is the class imbalance issue, where the data are not uniformly distributed among classes. As humans are not stressed for a large portion of their daily lives, stress datasets collected outside of a lab have a high chance of being imbalanced. One study considered both data from inside and outside a lab to compare how stress could be detected in these disparate conditions [6]. In our own analysis of real-world data, we found that performance was poor when no steps were taken to address the class imbalance, so we had to address this issue. Recent work investigated various methods of mitigating imbalance in data, and created novel improvements to common strategies like undersampling and oversampling [7], [8], [9].

Like these prior works, we investigate the issue of automated stress detection from wearable sensor data. Unlike many of these studies, which focus on trying to find the best ways to detect stress from wearable data, we focus on the performance of detecting stress with fewer modalities. Because the data classes are imbalanced, we compare some of the well-known methods to resolve this issue as well. Our goal is to make stress detection more accessible and thus investigate how well we can detect stress with fewer sensor requirements and computational cost.

III. Datasets

Our study analyzed stress detection based on two datasets, WESAD and Alcohol and Drug Abuse Research Program (ADARP). The WESAD dataset [10] was gathered in a lab setting. Subjects were put in several different situations to prompt them into stressful states. These situations were intermixed with other scripted behaviors. The ADARP data was collected from individuals suffering from alcohol use disorder outside of a lab setting, where stress was recorded on an Empatica E4 wearable band by the subject logging when they felt stress throughout the day (by tapping the push button on E4).

The WESAD dataset was gathered for the purpose of improving stress classification. Data were gathered from a total of 17 subjects. The subjects were graduate students at the institution of the researchers who created the data. Each subject was equipped with several sensors placed on the chest as well as on Empatica E4 wearable devices. The sensors from two of these subjects malfunctioned and were unusable, so only 15 of the subjects' data are available. From these 15 subjects, the data from both the chest and wearable device sensors are available, but our testing focused only on the data from the wearable device sensors. The wearable device sensor modalities included accelerometer, electrodermal activity, blood volume pulse, and temperature. The subjects were put through several different conditions: baseline, amusement, stress, and meditation. Baseline data were gathered for 20 minutes after the sensors were equipped and subjects were sitting or standing at tables with magazines provided for neutral reading to induce a neutral state. The amusement label data were gathered while subjects watched several funny videos. Stress conditions were created through the Trier Social Stress Test [11], consisting of a public speaking task and a mental arithmetic task. These tests were performed in front of a three-person panel of human-recourse specialists. The students were told to make a good impression on the panel to boost their career options (later the subjects were informed that the panel members were actually just researchers). The meditation condition was created by putting the subjects through a guided meditation. The meditation condition was initiated after both the amusement and stress conditions to calm the subjects. In our testing, we utilized the meditation, baseline, and stress conditions, to create stressed and non-stressed classes for training the learning algorithms. We used this dataset to evaluate whether single modalities would have comparable performance to multiple modalities using several different learning algorithms.

The ADARP dataset [12] was collected with the goal of investigating the relationships between sensor data and self-reported stress. Data were gathered from 11 subjects, 10 of whom were female. The subjects were adults seeking help at a facility for mental health. Inclusion criteria were that subjects must be age 18-65 years and have self-reported consuming four or more standard drinks (drinks containing roughly 14 grams of pure alcohol) in a single day 5 or more times in the previous 60 days. Data were gathered from three sources: a daily diary of self-reported emotions, cravings, and stress (four times a day), sensor data from Empatica E4 wearable devices, and structured daily interviews for qualitative evaluation of alcohol use. This study produced data from the same four sensor modalities included in the WESAD dataset. We used these data in addition to the WESAD data to determine if single modalities could perform comparably to a combination of multiple modalities. Additionally, we want to determine whether predictive performance from lab-based data was comparable to data collected outside the lab.

IV. Methods

To compare the four different sensor modalities available on the E4 wearable devices, acceleration, blood volume pulse, electrodermal activity, and temperature, we ran several tests on each modality from both datasets to determine which would yield the highest performance. To begin this process, we first preprocessed the datasets to make them compatible with our learning algorithms1. To format the WESAD dataset, we began by first separating out the wearable device data, as the chest sensor data are not relevant for our purposes. From the wearable device data, we removed data for the class labels that were not defined, should be ignored, or were categorized as "amusement." With the remaining labels, we combined "baseline" and "meditation" categories to form a "nonstress" class and categorized the remaining data as the "stress" class. The changes to the label distribution can be seen in Figures 1

[1] All code for formatting, training algorithms, and evaluation can be found online at https://github.com/RyanCHolder/Comparing-Single-Modalities.

and 2. We partitioned the remaining data into one-second windows with 50% overlap. The number of data instances in each window varied across modalities, as each modality was sampled at a different frequency (32 Hz for acceleration, 64 Hz for blood volume pulse, 4 Hz for electrodermal activity, and 4 Hz for temperature). These windows were used as the training data for our learning algorithms.

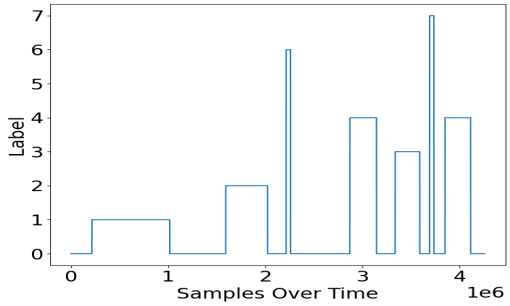

Fig. 1. WESAD label distribution for single subject before preprocessing.

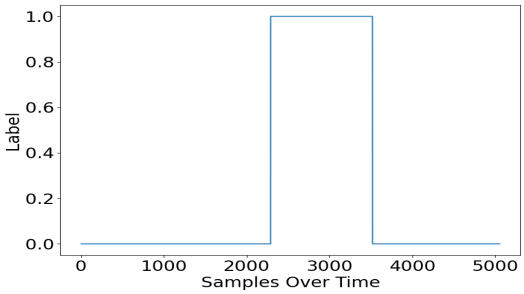

Fig. 2. WESAD label distribution for single subject after preprocessing.

To format the ADARP data, we defined instances of the stress class to encompass all data occurring within 20 minutes (10 minutes in each direction) of the subject's stress tags. All other nonstress data within an hour in each direction of that tag was removed, because the subject's stress state during that period is unknown. The labeled data were formed into one-minute windows with 50% overlap, creating the ADARP data that we used for training and testing the classifiers. The label distribution for a single subject from the ADARP dataset can be seen in Figure 3.

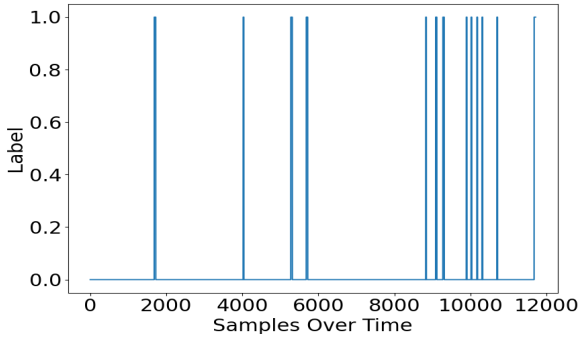

Fig. 3. ADARP label distribution for single subject.

An additional dataset was created for both WESAD and ADARP by extracting statistical features that describe each window of data. The statistical features were mean, median, minimum, maximum, standard deviation, skew, kurtosis, and interquartile range. We initially performed separate experiments with the raw data and the statistical features. We report results for only the raw data with the longer tests, as the performance using the statistical features was very similar to the performance using the raw data. These formatted datasets were used to train our learning algorithms.

To classify stress in the datasets we compare three learning algorithms: k-nearest neighbors (KNN), decision tree (DT), and a convolutional neural network (CNN). The k-nearest neighbors classifier was trained with uniform weights, a Minkowski distance metric, and Euclidean distance for the Minkowski power parameter. The decision tree classifier was trained with Gini to measure split quality, choosing the best option at each split, no maximum depth (expanding until all leaves are pure), a minimum of two samples to split a node, one minimum sample to be considered a leaf node, an unlimited number of allowable leaf nodes, a minimum impurity decrease of zero, and a minimum impurity split of zero.

The convolutional neural network consisted of two 2D convolutional layers, the first containing 10 filters and the second 20, both with 'same' padding. The CNN was trained using ReLU (rectified linear unit) activation and a kernel size of five. The convolutional layers were followed by a singular unit dense layer with sigmoid activation for producing output. The convolutional model was compiled with binary cross-entropy as the loss function, Adam as the optimizer, and accuracy as the training metric. The model was trained with 10 epochs, a batch size of 30, and a 20% validation split. For learning on WESAD, the algorithms were trained with no class weighting. However, in resolving the class imbalance issue with the ADARP dataset, class weighting was used that we will describe later.

V. Experimental Conditions

To gather results for each modality, we ran two varieties of tests. The first variety was a randomly selected train-test split from the combined data of all subjects. The split was chosen with 75% training data and 25% testing data. We chose this split to ensure plenty of data for training, and a portion of testing data large enough to avoid a biased sample. In our testing with this split we found algorithms to converge consistently on training data, as well as perform comparably on testing data. Each algorithm was trained with the training portion of the data for each modality individually for a total of 10 epochs. The trained models were used to predict class labels for the remaining 25% of the data, and these predictions were evaluated based on both accuracy and f1-score. The results for each iteration were averaged to produce an accuracy and f1-score for each modality and each learning algorithm.

The second experiment we ran was a leave-one-subject-out test. For this experiment, we combined the data of all but one subject to form the training data and used the data of the left out subject as the testing set. This was run for one iteration per subject, where each subject was used as testing once. Each modality was trained individually on each algorithm for every iteration, and the resulting models were used to predict class labels for the remaining subject. Accuracy and f1-score were used to evaluate the predictions, and the average of these results for each iteration were computed to yield one accuracy and one f1-score value for each modality using each algorithm.

These two varieties of experiments were run on both datasets; however, the balance between nonstress and stress classes in each dataset was very different. After our formatting, the WESAD exhibited a stress to nonstress ratio of approximately 1:3, while the ADARP dataset had a ratio of roughly 1:16. Because of the large imbalance in the ADARP data, we also added four class imbalance solutions to the above testing methods and tested those on the ADARP data. The solutions we used were majority class undersampling, minority class oversampling, class weighting, and a combination of the three methods. In the case of undersampling, we randomly selected a portion of nonstress data to include that was equivalent to the amount of stress data for each training set. In the case of oversampling, we used the Imbalanced-learn library's Synthetic Minority Over-sampling Technique (SMOTE) [13] algorithm to generate synthetic data to create an equal amount of stress and nonstress data. For our weighting method, we set the class weights when training to 16 on the stress class, and 1 on the nonstress class. For our combination method, we performed undersampling in the same way as above, while selecting a portion of nonstress that was ten times the size of the stress class. We then performed oversampling on the resulting set to bring the stress class up to one-fourth the size of the nonstress class. With the new dataset, the learning algorithms were trained with the class weights set at four for the stress class and one for the nonstress class. In both the standalone weighting solution and the combined method, we only used the decision tree and convolutional neural network algorithms, as k-nearest neighbors does not support class weights as a parameter for its learning. We tested each of these class imbalance solutions using the same two testing varieties described above, using the adjusted dataset where applicable in place of the original data in the random sampled test, and creating the adjusted dataset at the beginning of each iteration for the leave-one-out test.

To create a baseline of comparison for our results on singular modalities, we also ran the same tests using a combination of all four modalities. To combine the modalities, we changed some of the data preprocessing to downsample each modality as needed to fit the same number of instances per window, which was necessary because of the differing sampling frequencies. This was performed by selecting the first value of every quarter of a second from each modality, resulting in no change in the electrodermal activity or temperature data, but reducing the quantity of data from both the accelerometer and blood volume pulse sensor values. We combined the resulting downsampled data into a singular dataset by concatenating all the sensor values together in each instance within each window. The resulting dataset was then run with the same random sample test, as well as the leave-one-out test. Because of the class imbalance in the ADARP data, we performed these tests with the same oversampling we performed on the singular modalities when testing combined modalities on the ADARP data, generating enough synthetic data to create equal proportions of stress and nonstress data. We choose to use oversampling rather than one of the other imbalance solutions when testing with all modalities because it was the best-performing solution we tested on singular modalities with the ADARP data.

VI. Results

We can see that singular modalities showed promising performance on the WESAD dataset. The results from the statistical and raw features were very similar, so we focus on just the results of the raw data. While not all modalities performed close to the combined modalities, we can see in the random holdout validation that acceleration only decreased accuracy by 0.0493 and f1-score by 0.1019 on average across all algorithms (see Figure 4). Electrodermal activity closely followed acceleration in the random holdout validation, with an accuracy decrease of 0.0689 and an f1-score decrease of 0.1387 from the combined modalities on average across all algorithms (see Figure 4). Similarly, on the leave-one-out validation, we saw electrodermal activity perform very strongly, outperforming the combined modalities by 0.0511 in accuracy, and 0.0950 in f1-score on average across all algorithms (see Figure 5). We suspect that the combined modalities did not perform as well as electrodermal activity due to the decrease in training data because of downsampling, as well as an increase in features without a corresponding increase in data volume, both contributing to underfitting. Despite the potential of underfitting, we can clearly see the potential of singular modalities to yield performance similar to combined modalities.

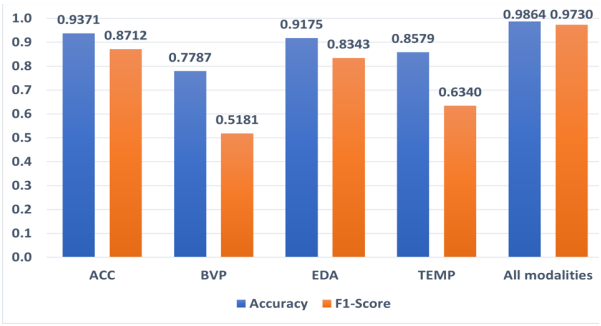

Fig. 4. Average performance of 3/4-1/4 validation on WESAD.

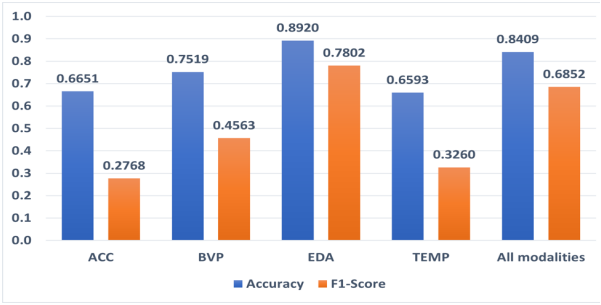

Fig. 5. Average performance of leave-one-subject-out validation on WESAD.

From our testing on the ADARP dataset we saw results that also suggest singular modalities have potential to give performance similar to that of multiple combined modalities. Our results from oversampling were best out of all our class imbalance solutions, so we will focus on those results. In our random holdout testing, we found that all modalities performed similarly, and on average were short of the performance of the combined modalities by 0.0446 in accuracy and 0.0373 in f1-score across all algorithms (see Figure 6). This result shows that in the dataset with significantly more data available (though still

very imbalanced) singular modalities continue to show performance comparable to that of the combined modalities. Our leave-one-subject-out results were quite extreme, as on average the singular modalities outperformed the combined modalities by 0.2128 in accuracy and were short from the combined modalities by 0.2525 in f1-score (see Figure 7). These results show that our singular-modality algorithms were overfitting much more than our combined modality algorithms, however the high accuracies still show potential for accurately detecting stress states from singular modalities. We also suspect that the combined modalities may have performed worse in this case for similar reasons as in the WESAD testing, as downsampling to align the modalities greatly reduced the amount of training data and increasing the number of features without increasing the amount of data leads to underfitting. Summaries of the experimental results are provided in Tables 1 through 4.

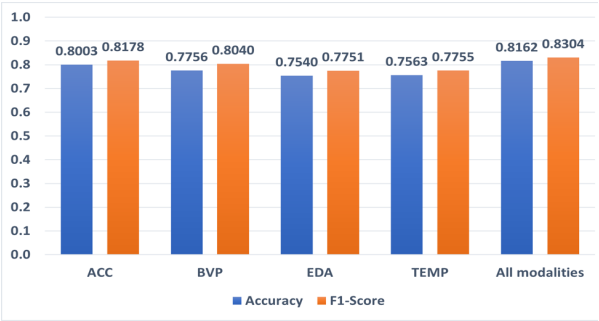

Fig. 6. Performance of 3/4-1/4 validation on ADARP with oversampling.

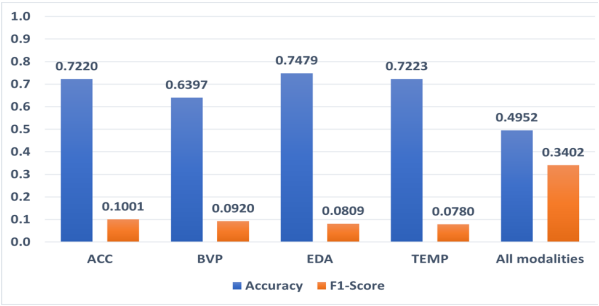

Fig. 7. Performance of leave-one-subject-out validation on ADARP with oversampling.

TABLE I. 3/4-1/4 VALIDATION WITH WESAD DATASET.

		Raw Data		Statistical Features	
	Modality	***Accuracy***	***F1-Score***	***Accuracy***	***F1-Score***
K-Nearest Neighbors	ACC	**0.9587**	**0.9146**	**0.9566**	**0.9140**
	BVP	0.7974	0.5423	0.7582	0.4637
	EDA	0.9199	0.8394	0.8313	0.6474
	TEMP	0.8940	0.7865	0.7857	0.5414
	All	0.9871	0.9743		
Decision Tree	ACC	**0.9427**	**0.8860**	**0.9479**	**0.8967**
	BVP	0.7513	0.5102	0.7186	0.4481
	EDA	0.9115	0.8248	0.8500	0.7012
	TEMP	0.9212	0.8389	0.8094	0.5684
	All	0.9886	0.9775		
CNN	ACC	0.9099	0.8129		
	BVP	0.7872	0.5018		
	EDA	**0.9209**	**0.8388**		
	TEMP	0.7584	0.2766		
	All	0.9835	0.9673		

TABLE II. LEAVE-ONE-SUBJECT-OUT VALIDATION WITH WESAD DATASET.

		Raw Data		Statistical Features	
	Modality	***Accuracy***	***F1-Score***	***Accuracy***	***F1-Score***
K-Nearest Neighbors	ACC	0.6528	0.2566	0.7331	**0.4763**
	BVP	0.7713	0.4831	**0.7377**	0.4177
	EDA	**0.8975**	**0.7885**	0.6815	0.3421
	TEMP	0.6463	0.3493	0.6633	0.3456
	All	0.8391	0.6812		
Decision Tree	ACC	0.6668	0.3685	**0.7071**	**0.4164**
	BVP	0.7271	0.4732	0.6947	0.4099
	EDA	**0.8618**	**0.7293**	0.6912	0.3976
	TEMP	0.6522	0.3621	0.6834	0.3316
	All	0.8508	0.6992		
CNN	ACC	0.6758	0.2052		
	BVP	0.7573	0.4127		
	EDA	**0.9167**	**0.8229**		
	TEMP	0.6794	0.2667		
	All	0.8328	0.6753		

TABLE III. CLASS IMBALANCE STRATEGIES USING ADARP DATASET WITH 3/4-1/5 VALIDATION.

		Under-sampling		Over-sampling		Weighting		Combination	
	Modality	***Accuracy***	***F1-Score***	***Accuracy***	***F1-Score***	***Accuracy***	***F1-Score***	***Accuracy***	***F1-Score***
K-Nearest Neighbors	ACC	**0.6119**	**0.5743**	0.8388	0.8607				
	BVP	0.4943	0.5098	0.7006	0.7698				
	EDA	0.5768	0.5388	**0.8424**	**0.8629**				
	TEMP	0.5751	0.5736	0.8208	0.8475				
	All			0.9043	0.9125				
Decision Tree	ACC	**0.6067**	**0.6039**	0.8437	0.8495	**0.8918**	**0.1631**	0.8125	0.5454
	BVP	0.5142	0.5130	0.7740	0.7847	0.8830	0.0701	0.7659	0.4337
	EDA	0.5704	0.5676	0.8339	0.8395	0.8880	0.1191	0.8006	0.5176
	TEMP	0.5718	0.5714	**0.8681**	**0.8721**	0.8898	0.1162	**0.8150**	**0.5521**
	All			0.8757	0.8795				
CNN	ACC	0.5473	0.5364	0.7185	0.7431	0.3705	**0.1224**	0.5457	0.4029
	BVP	0.5108	0.5088	**0.8524**	**0.8574**	**0.7066**	0.0987	**0.7602**	**0.5156**
	EDA	0.5409	0.5334	0.5857	0.6229	0.4481	0.1171	0.5149	0.3442
	TEMP	**0.5803**	**0.5812**	0.5800	0.6069	0.1758	0.1068	0.4461	0.3329
	All			0.6685	0.6992				

TABLE IV. CLASS IMBALANCE STRATEGIES USING ADARP DATASET WITH LEAVE-ONE-SUBJECT-OUT VALIDATION.

		Under-sampling		Over-sampling		Weighting		Combination	
	Modality	***Accuracy***	***F1-Score***	***Accuracy***	***F1-Score***	***Accuracy***	***F1-Score***	***Accuracy***	***F1-Score***
K-Nearest Neighbors	ACC	**0.6536**	**0.1140**	**0.6625**	**0.1080**				
	BVP	0.4658	0.1075	0.4348	0.1066				
	EDA	0.5547	0.1043	0.6608	0.0991				
	TEMP	0.5218	0.1097	0.6452	0.1028				
	All			0.4852	0.2571				
Decision Tree	ACC	**0.5631**	**0.1185**	**0.7066**	**0.1088**	0.8588	**0.0662**	0.7641	0.0854
	BVP	0.4900	0.1088	0.6909	0.0922	**0.8742**	0.0603	**0.7911**	0.0796
	EDA	0.5043	0.1073	0.6898	0.0902	0.8388	0.0545	0.7675	0.0792
	TEMP	0.4965	0.1054	0.6318	0.0987	0.8652	0.0609	0.7729	**0.0887**
	All			0.5010	0.2740				
CNN	ACC	0.6441	0.0977	0.7970	**0.0835**	**0.6780**	**0.0915**	0.9349	0.0000
	BVP	0.6473	**0.1114**	0.7933	0.0771	0.6410	0.0853	0.9349	0.0000
	EDA	**0.8738**	0.0768	0.8931	0.0533	0.5129	0.0856	0.9349	0.0000
	TEMP	0.8683	0.0537	**0.8899**	0.0324	0.3482	0.0715	0.9349	0.0000
	All			0.4994	0.4896				

VII. Conclusions and Future Work

The goal of our work was to investigate the predictive performance of individual sensor modalities in detecting stress states using machine learning techniques. To accomplish this, we compared performance of individual sensor modalities from two stress detection datasets, using several different machine learning techniques. Our results lead us to believe that there is potential for singular modalities to perform comparably to multiple modalities. Including more features and data available will certainly lead to higher performance; however, the performance in many cases of singular modalities were close to that of the combined modalities. We believe that our results provide evidence to support the hypothesis that singular modalities can predict stress at a high enough level to be used in place of multiple modalities to save on computational requirements.

While our work gives some preliminary evidence for the promise of reducing sensor modalities to save computational requirements to detect stress, there is still a lot more work that is necessary to refine the use of singular modalities in stress detection. Our work can be enhanced by testing a greater variety of algorithms. In particular, the structure of the convolutional neural network may be refined to improve performance, as these networks show promise of being the most powerful tool for processing time series data. Future work could test several convolutional architectures, as well as increasing depth and number of epochs to optimize the performance of singular modalities and compare to the performance of combined modalities to see if the results are comparable in a more optimized environment.

Future work can also be directed toward resolving the class imbalance problem that is common in stress detection datasets. Our oversampling solution did yield an increase in performance compared to learning from imbalanced data, but we believe that more can be done to refine how the class imbalance is addressed to give increased performance of both combined and singular modalities. We believe the class imbalance may have influenced our results, especially in our experiments that utilized the ADARP dataset. As class imbalance solutions improve, employing these solutions to adjust imbalanced datasets like the ADARP dataset can lead to results with less irregularity to better show how singular modalities perform against multiple modalities.

VIII. References

[1] E. Sagbas, S. Korukoglu, and S. Balli, "Stress Detection via Keyboard Typing Behaviors by Using Smartphone Sensors and Machine Learning Techniques," *J. Med. Syst.*, vol. 44, no. 68, 2020.

[2] 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.

[3] K. Kyriakou *et al.*, "Detecting Moments of Stress from Measurements of Wearable Physiological Sensors," *Sensors*, vol. 19, no. 17, 2019.

[4] S. Aqajari, E. Naeini, M. Mehrabadi, S. Labbaf, A. Rahmani, and N. Dutt, "GSR Analysis for Stress: Development and Validation of an Open Source Tool for Noisy Naturalistic GSR Data." 2020.

[5] C.-P. Hsieh, Y.-T. Chen, W.-K. Beh, and A.-Y. Wu, "Feature Selection Framework for XGBoost Based on Electrodermal Activity in Stress Detection," 2019.

[6] M. Gjoreski, H. Gjoreski, M. Lustrek, and M. Gams, "Continuous stress detection using a wrist device: in laboratory and real life," in *2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016, pp. 1185–1193.

[7] X. W. Liang, A. P. Jiang, T. Li, Y. Y. Xue, and G. T. Wang, "LR-SMOTE — An improved unbalanced data set oversampling based on K-means and SVM," *Knowledge-Based Syst.*, vol. 196, p. 105845, May 2020, doi: 10.1016/J.KNOSYS.2020.105845.

[8] A. Guzmán-Ponce, J. S. Sánchez, R. M. Valdovinos, and J. R. Marcial-Romero, "DBIG-US: A two-stage under-sampling algorithm to face the class imbalance problem," *Expert Syst. Appl.*, vol. 168, p. 114301, Apr. 2021, doi: 10.1016/J.ESWA.2020.114301.

[9] A. Bria, C. Marrocco, and F. Tortorella, "Addressing class imbalance in deep learning for small lesion detection on medical images," *Comput. Biol. Med.*, vol. 120, p. 103735, May 2020, doi: 10.1016/J.COMPBIOMED.2020.103735.

[10] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Laerhoven, "WESAD: Multimodal Dataset for Wearable Stress and Affect Detection," 2018.

[11] C. Kirschbaum, K. Pirk, and D. Hellhammer, "The 'Trier Social Stress Test' – A Tool for Investigating Psychobiological Stress Responses in a Laboratory Setting," *Neuropsychobiology*, vol. 28, no. 1–2, 1993.

[12] P. Alinia *et al.*, "Associations Between Physiological Signals Captured Using Wearable Sensors and Self-reported Outcomes Among Adults in Alcohol Use Disorder Recovery: Development and Usability Study," *JMIR Form. Res.*, vol. 5, no. 7, 2021.

[13] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.