Experimental Protocol for Detection of Stress using Photoplethysmography Signals

Vishnu Kannan
vishnukanan@gmail.com

Ketki Ambekar
ambekar.ketki@gmail.com

Vaidic Joshi
er.vaidic@gmail.com

Pradeep Ramadasan
pradeep.ramadasan@gmail.com

1 Hypothesis

Problem Definition: We're building a predictive model to detect stress (binary classification) from Blood volume pulse signals obtained from photoplethysmogram (PPG) devices commonly found on fitness and clinal devices including Fitbit, Apple Watch, etc. Respiratory rate (RR), often referred to as breathing rate, is the number of breaths a person takes per minute. A normal resting RR for adults ranges from 12 to 20 . Abnormal changes in respiratory rate are an accurate indicator of physiological conditions such as anxiety, hypoxia, hypercapnia, metabolic and respiratory acidosis.
A diverse body of research studies has indicated the significance of respiration rate for forecasting events such as cardiac arrest, patient deterioration, and care escalation. RR's reliable measurement devices are bulky and cumbersome, and are mainly used for patients in hospitals. With the development of wearable technologies, any change in a subject's physiological functional stage can monitored in an everyday setting, by using PPG. PPG signals can easily and cheaply be collected continuously and noninvasively using a wide range of inexpensive, convenient, and portable wearable devices (.e.g., smart watches, rings, etc.).

1.1 Why is the problem important and why is it hard?

Stress and Anxiety can lead to both short term and long term negative physiological changes [9]. More recently covid-19 has been studied to trigger depression, anxiety and other neurological conditions [3]. Being able to detect ongoing stress or anxiety condition and empowering humans to take proactive steps to help themselves would be invaluable to our society at large. PPG as a signal is prone to motion artifacts which makes the signal vary a lot in amplitude and also carry noise. Hence modeling it isn't simple.
Past research has been centered around using time domain features like Inter beat intervals, Heart Rate, and more statistics around them, or frequency domain features like power spectrum [7], [6]. However estimating these features is difficult and error prone where signal quality greatly determines ability to detect peaks which form the basis of many time domain features. Complex signal processing is often needed and the ability to generalize has also not been demonstrated. Recent research with deep convolutional neural networks [6], [5] have shown the ability to model time domain PPG signals without requiring any feature extraction and detect heart rate and stress directly. The flip side of deep neural networks is that they are computationally infeasible for low power wearables. Being able to develop powerful models that can generalize well and be computationally cheap is still an open research problem.

1.2 Key limitations

Classical ML techniques require extensive domain expertise to design features suitable for a comprehensive representation of PPG and the detection of class-differentiating patterns. Features commonly extracted from PPG time series are morphological descriptors, time domain statistics, frequency domain statistics, nonlinear measures, wavelet based measures, and cross-correlation measures. These signal processing techniques are also not guaranteed to work all the time and hence require clever hacks in the algorithm to work around corner cases.

Training a DL model from scratch also requires a large amount of labeled training data and is a major constraint in biomedical applications due to the limited amount of labeled data. A possible solution to overcome this limitation is to use representation learning where similar to NLP,

representations learned to solve multiple standard tasks could help build models that can identify several other conditions detectable from PPG signals including stress.

1.3 What are the opportunities to improve upon existing work and possible experiments to explore

Deep convolutional networks have recently emerged as a powerful method for the detection of abnormalities in physiological signals, encouraging applications of PPG. Representation learning using auto encoders are also shown to be promising.
Unlike ML, deep learning models automatically learn feature representations, sparing the tedious task of feature crafting.

2 Data

We are working with three datasets:

2.1 WESAD:

Link::https://ubicomp.eti.uni-siegen.de/home/datasets/icmi18
This dataset records Blood Volume Pulse (BVP), Electro Dermal Activity (EDA), and some other features collected from PPG devices called RESPIBAN (from chest) and Empatica (from wrist) of 17 participating subjects.

This multimodal dataset features physiological and motion data, recorded from both a wrist- and a chest-worn device, of 17 subjects during lab studies. The following modalities are included in the dataset: blood volume pulse, electrocardiogram, electrodermal activity, electromyogram, respiration, body temperature, and three-axis acceleration.This dataset bridges the gap between previous studies on stress and emotions, by containing three different affective states (neutral, stress, amusement). In addition, self-reports of the subjects, which were obtained using several established questionnaires, are contained in the dataset. Furthermore, a benchmark is created on the dataset, using well-known features and standard machine learning methods. Considering the three-class classification problem (baseline vs. stress vs. amusement), the achieved classification accuracies of up to 80%. In the binary case (stress vs. non-stress), accuracies of up to 93% were reached. Note, the dataset authors did not share any other metrics like AUC-ROC, F1 score, log likelihood, etc.

2.2 PPG-DALIA:

Link:https://archive.ics.uci.edu/ml/datasets/PPG-DaLiA
This dataset is created by creators of WESAD. It is similar in structure to the previous one with data from 15 subjects. The difference being, while the activities in WESAD were designed to be stress inducing, DALIA focused on normal daily activities like walking, sitting, biking, etc., and aims to help estimate heart rate from PPG signals. It is useful for this project to help learn representations using deep neural networks.

2.3 UBFC

Link:https://ieee-dataport.org/open-access/ubfc-phys-2
The UBFC Dataset monitors Blood Volume Pulse (BVP) and Electrodermal Activity (EDA) of 56 participants in three tasks. The tasks are designed to be anxiety inducing. The participants are asked to self-score their anxiety response before and after these tasks. We only used the BVP data from the original dataset, processed it to a suitable format.

2.4 Preprocessing employed in this project

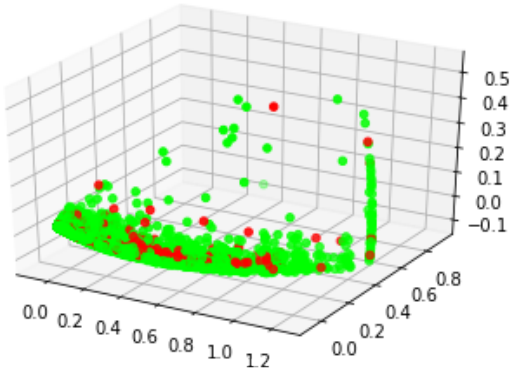
For the purposes of stress detection, only BVP (blood volume pulse) signals are being employed. This is done such that the models being build can be extended to work with remote PPG sensing (rPPG) via video cameras in the future. In addition just the PPG signal has been shown to detect conditional like Atrial Fibrilation [8]. The common windowing technique of 8 second duration with 6 second overlap has been employed [6] in this project as well.
The final dataset format will include bvp signals sampled at 64 HZ where each sample window contains 512 measurements. The stress labels for the Wesad [7] dataset provided was sampled at 700 Hz to match the ECG signal frequency. We downsampled that to match the PPG signal frequency of 64Hz.

With the Deep-PPG dataset [6], windows were already provided with heart rate labels provided for each window. This data has been employed as-is.

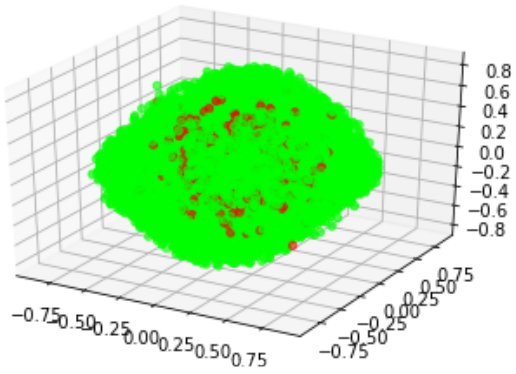
With the UBFC-PHYS dataset [4] no stress labels were provided. Instead task description and participant self-assessment of stress affects measured via standard psychological tests were provided. We are left to choose to either go with the

self-assessment to identify personalized per-task stress labels for participants or stick to the previous dataset authors where they assumed the tasks chosen to induce stress would have induced stress in all participants. Either path carries the risk of bias and we chose to go with self-assessment.

To get a sense of the data distribution with the WeSAD dataset, we ran PCA and tried to look at the distribution of stress vs non-stress samples. We visualized both the raw PPG signal windows and also PPG specific signals extracted by an open source library [1]. As shown below in figure 'a' the heartpy features seem to provide some separation of the classes for building a linear classifier. However as shown in figure 'b', the raw time domain signals do not seem readily separable.



(a) Plot of PCA with 3 dimensions using heartpy derived time domain features



(b) Plot of PCA with 3 dimensions using raw time domain data

Figure 1: Principle Component Analysis

3 Evaluation and Metrics

The WESAD dataset is highly imbalanced, because during the study protocol, only a small subset of the duration were associated with stress inducing activities. The dataset has about 11% of the samples belonging to stressed state. This makes usage of model metrics like Accuracy or log likelihood or F1 score not ideal. Instead we chose to employ AUC-ROC as it captures both true positivity rate and false positivity rate where the latter should be low for this problem statement. Hence we seek a model that has high Specificity, and reasonable Sensitivity. ROC-AUC captures this intent well where a model with ROC-AUC closer to 1 would be what we'd choose. Cross-validation was applied with leave one group out strategy. Given the variability of PPG signals between humans, this cross validation strategy was identified as a way to discern ability of a model to generalize. The final accuracy is reported as the mean of all the testing accuracies where one participant is left out for testing and others for training in each iteration.

Before evaluating models, we randomly picked one participant an left their data as the test set. The rest of the participants (14 in total) were used for model selection with cross validation employed to pick a model.

Models being evaluated were training 14 times where each time one of the participants were used as the validation set for evaluating the model. Finally the average ROC AUC score was used to compare models.

Given that the models we trained thus far haven't shown ability to discern stress signals in the dataset we are yet to venture into evaluating bias/variance tradeoff and identify aspects like impact of sample size on our models for example.

4 Models

We first developed a baseline model which given the class imbalance will always predict not stressed (0 class label). This model will yield an accuracy of about 88%.

The following model kinds were evaluated: Logistic Regression using SGD, SVM with SGD, and SVM with a radial basis function kernel that had 100 dimensions. Raw time domain data was scaled using a technique that removes the median and scales the data based on IQR to work around outliers which were common in the dataset.

As expected in the PCA analysis, none of these linear models were able to perform better than the baseline model, ROC-AUC was around 0.5.

Before venturing into training deep neural networks, we tried to employ the features obtained

from heartpy library [**heartpy**] which runs peak detection algorithm to identify heart beat and computes several statistics based on the heart beat intervals.
An SVM trained with a RBF kernel showed some promise with a ROC-AUC score of 0.71 and accuracy of 0.87. Upon inspecting the confusion matrix, this model chose to classify all signals as belonging to non-stress class, thereby mimicking our baseline model.

5 General Reasoning

•Explain how the data and model(s) come together and inform your core hypothesis

Our core hypothesis of detecting stress from PPG signals was already shown to be possible in [7] and [5]. But we are yet to develop suitable features or model architectures that can match the accuracies of the models described in those publications.

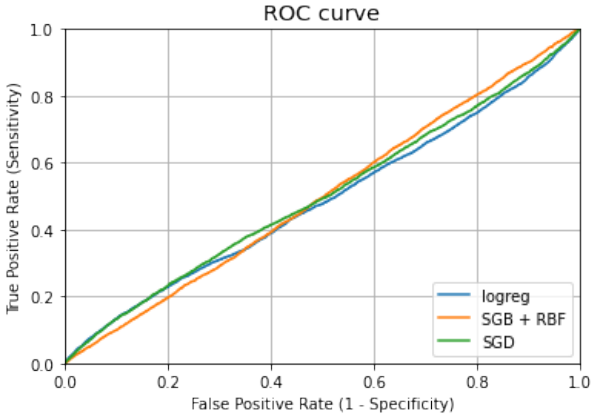


Figure 2: "Plot of ROC curve using raw time domain BVP signals"

As shown in Figure '2', standard classical ML models like SVM or Logistic regression aren't able to separate stress from non-stress bvp signals. We need to employ non-linear models.

6 Summary of Progress

1. We established a clear hypothesis that we intend to test - Psycological stress condition can be detected from PPG Blood volume pulse signals with high specificity and reasonable sensitivity.

2. We identified relevant datasets, studied the datasets, and pre-processed them to begin modelling.

3. We tried Principal Component Analysis to get a view into class separation within the dataset.

4. We identified techniques to compute common time domain features from BVP signals.

5. We identified ROC-AUC as the appropriate metric to employ for this binary classification problem.

6. We established leave one group out cross validation as the preferred cross validation techniques to evaluate models.

7. We established a dummy baseline model which can achieve a accuracy of 87%.

8. We trained a few standard classical ML models like Logistic Regression and SVMs to get a sense of linear separability of the dataset which was shown to be difficult in PCA analysis and reinforced our belief that raw time domain signals will require complex model architectures to extract relevant features.

9. We employed time domain features like heart rate, inter beat internal, etc. (a total of 13 features) and that resulted in a SVM classifier that matches that of the dummy classifier.

10. We identified several next steps to pursue including solving data imbalance problem through techniques like SMOTE, employing deep neural networks, improving time domain and trying out frequency domain features.

6.1 Next Steps:

Our dataset is highly imbalanced towards non-stressed readings than it is towards stressed readings. We are exploring techniques to balance it out, which include undersampling larger labels, oversampling scarce labels, and using data augmentation techniques such as SMOTE, and time series specific data sampling techniques as mentioned in [2]

We also intend to experiment with deep neural networks, where we will start with reproducing the auto encoder that employed simple CNNs in [5] followed by expanding the model architecture to include ResNext modules which were shown to be useful in [8]. With the learned representations, we will attempt to build a classifier for two tasks - heart rate detection and stress prediction to see how well the representations can generalize.

We will attempt to spend some time improving our time domain feature extraction algorithms and trying out frequency domain features in an attempt to match the benchmark results published in the original dataset papers.

References

[1] In: (). URL: https://python-heart-rate-analysis-toolkit.readthedocs.io/en/latest/.

[2] Qingsong et al. "Time Series Data Augmentation for Deep Learning: A Survey". In: (2021). URL: https://arxiv.org/abs/2002.12478.

[3] Mario Gennaro Mazza et al. "Anxiety and depression in COVID-19 survivors: Role of inflammatory and clinical predictors". In: *Brain, behavior, and immunity* 89 (2020), pp. 594–600.

[4] Rita Meziatisabour et al. "UBFC-Phys: A Multimodal Database For Psychophysiological Studies Of Social Stress". In: *IEEE Transactions on Affective Computing* (2021).

[5] Ali Oskooei et al. "Destress: deep learning for unsupervised identification of mental stress in firefighters from heart-rate variability (hrv) data". In: *Explainable AI in Healthcare and Medicine*. Springer, 2021, pp. 93–105.

[6] Attila Reiss et al. "Deep ppg: Large-scale heart rate estimation with convolutional neural networks". In: *Sensors* 19.14 (2019), p. 3079.

[7] 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.

[8] Yichen Shen et al. "Ambulatory atrial fibrillation monitoring using wearable photoplethysmography with deep learning". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 1909–1916.

[9] Habib Yaribeygi et al. "The impact of stress on body function: A review". In: *EXCLI journal* 16 (2017), p. 1057.