Wearable Stress and Affect Recognition

Samyak S Sarnayak
dept. Computer Science and Engineering
PES University
Bengaluru, India
samyakssarnayak@pesu.pes.edu

Pranav Kesavarapu
dept. Computer Science and Engineering
PES University
Bengaluru, India
pranavkesavarapu@gmail.com

Abhijit Mohanty
dept. Computer Science and Engineering
PES University
Bengaluru, India
mohantyabhijit074@gmail.com

Raghav Aggarwal
dept. Computer Science and Engineering
PES University
Bengaluru, India
raghavaggarwal03.ra@gmail.com

***Abstract*—Wearable devices provide a large amount of multi-modal data which can be used for affect recognition. One such dataset is the WESAD dataset consisting of sensor data with labels for the test subject's affect or state of mind. In this paper, we first review existing approaches of stress and affect recognition. We find from the data exploration that the data consists of highly varying time series with a very large number of data points. We do not find any significant correlation, missing values or inconsistencies. It is also found that the values differ significantly between test subjects which makes a generalized model difficult to build. We then perform dimensionality reduction using PCA and use the reduced dimensions to build clustering models. The clustering models do not perform well, but can identify stress from the sensor readings. Finally, we train classification models on data and get good accuracies and F1-score with most classification models - especially KNN and random forest.**

***Index Terms*—affect detection, stress, wearable, health, visualisation, exploratory data analysis, dimensionality reduction, clustering, classification, multi modal**

I. Introduction

Affect is defined as the experience of emotion [1]. Affect recognition is the prediction of affective state of a person (test subject) based on certain observable factors such as ECG, heart rate, etc. Affect recognition forms the building blocks of *Affective computing* which involves emotionally intelligent machines that can recognise and simulate emotions. Affective computing could potentially give rise to machines such as personalized tutors which take into consideration the state of mind of the student, a tool that lets a teacher know the activity/interest of students in the classroom or simply a machine that recognises when a person is under stress and recommends techniques to reduce them.

Stress is one of the major issues plaguing modern society. Stress could be easy to recognise for a person in certain conditions such as before an important deadline. Though, stress can have severe side-effects in the long term such as headaches, troubled sleeping or even cardiovascular diseases [2]. Stress detection could help people become aware of when they are in stress and help control it. This prevents stress from being going unnoticed and avoids long-term effects.

Many techniques have been explored for affect recognition. Some of the techniques where a single factor was used to detect affect include analysis of a person's face pictures, speech pattern and body language [8]. These are called uni-modal methods since they utilise a single factor or variable.

Multi-modal techniques are methods where multiple modalities of affect recognition are explored i.e., multiple factors are taken into account. Multi-modal data provides a large amount of data, both horizontally and vertically, which avoids relying on a single erroneous sensor data. According to D'Mello et al. [9], multi-modal data are usually around 10% better than similarly produced uni-modal data.

A good source of multi-modal affect-related dataset are **wearable devices** since they can record both physiological and inertial parameters. They provide rich data and have a easy-to-use form factor which allows more test subjects to be taken into consideration. Wearable devices are also an ideal platform for end user systems designed from affect recognition.

Thus, in this paper we look at Affect and Stress recognition using data from wearable devices. In particular, the *Wearable Stress and Affect recognition Dataset (WESAD)* is explored, visualised and conclusions are drawn from it.

II. Literature Survey

A. *Wearable affect and stress recognition: A review [2]*

Due to their good functionality and their small form factor, wearable systems are ideal for performing affect recognition. This Paper provided a clear understanding of the theoretical background. In this paper Schmidt et al. [2] showed different related works and the results regarding the wearable systems and stress recognition.

*1) **Assumptions**:* In order to get a high quality data in affect detection, strict and care full study protocols are required.

- The subject has not consumed any tobacco or drug before conducting the test. (In WESAD dataset)
- All the sensors are transmitting and receiving data accurately (WESAD dataset).

- Data from each test subject was assumed to be independent of any other subject.
- Semester Exams were approaching during the study of studentLife dataset More number of Female subjects were there in Different datasets such as (Eight Emotions, MAHNOB-HCI by Soleymani et al. [7]).

*2) **The major contribution of the paper**:* This Paper helps in providing Comprehensive comparison of the analysed wearable affect and stress recognition by using different classification models which were done by different authors. They have shown how different classification models such as KNN, SVM, ANOVA, LDA, ANFIS (adaptive neuro-fuzzy inference system), Decision Trees etc. and their results (the accuracy each model has given). Based on the comparison they came up with some prominent results which are:

- k-fold Cross-Validation (CV) was used for building the models
- They were able to show a clear distinction between field study and lab study by comparing the accuracy's.
- Deep Neural networks worked well for the datasets tested by Fernández-Delgado et al. [6]
- AdaBoost was applied by Mozos et al. [3] reaching an accuracy upto 94%.

*3) **Shortcomings and limitations**:* The above paper provided an overview of the theoretical background and they were able to provide comprehensive comparison of the analysed wearable affect and stress recognition but they were not able to show any form of Exploratory Data Analysis and any machine learning algorithm in use. They combined different works and data sets and how different authors have came up with different findings. Moreover they were unable to show any statistical findings rather they just compared machine learning techniques used by different authors.

*B. **Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection** [4]*

From the study it was found that stress lead to more than 30% of illnesses related to work [5]. The side-effects of stress makes it a priority to automate stress detection methods. It is a difficult task to differentiate stress and other emotions.

*1) **Assumptions**:* In order to get reliable and correct data subjects were asked to follow the following guidelines:-

- The subjects were restricted to consume any mood stabilizing drugs before conducting the experiment.
- The participants were forbidden from consuming caffeine and tobacco a few hours prior.
- The subjects were forbidden from doing heavy exercise on the day of the study.
- All the sensors are correctly working and properly connected to the subject with accurate precision.

*2) **The major contribution of the paper**:*

- An open dataset which is multi-modal also, is furnished. With the help of chest and wrist based sensors the data was recorded, each having high resolution sensor readings - *BVP, ECG, EDA, EMG, RESP, TEMP and ACC).*
- This new dataset considers three different affective states of brain, which are **baseline, stress, amusement**.
- Using different machine learning algorithms (Decision Tree, Random Fores, AdaBoost, Linear Discriminant Analysis and k-nearest neighbour), a baseline benchmark is obtained.

*3) **Shortcomings and limitations**:* In beginning of the experiment there were 17 test subjects but due to sensor malfunction for 2 subjects, sample of 15 test subject is available hence reducing the sample size of dataset. Moreover, sensors are prone to synchronization problem therefore, transmission and receiving delay may occur. Thus we consider each test subject is independent of other subject but it might not be the case.

*C. **Stress Detection from Multimodal Wearable Sensor Data***

In the earlier days, information about a person's state was retrieved from questionnaires and interviews. This was intrusive in nature and interrupting the task that is being carried out. The WESAD dataset contains data collected from non-intrusive wearable sensors. Using the sensory data, a personal stress detection system was created.

Indikawati et. al [10] implemented three classification algorithms , namely: Logistic Regression, Decision Tree and Random Forest. The classification conditions were baseline, stress, amusement and meditation. The model was trained on each subject (the dataset contained 15 subjects) since the physiological behaviour and changes vary from person to person.

The best and consistent personalized stress detection was using a Random Forest classifier with an accuracy of 88% - 99%. This was only with the data from the Empatica E4, a wrist-worn device. The questionnaires and the RespiBIAN, a chest-work device, were not included.

III. Data Exploration

In this paper we look at the Wearable Stress and Affect recognition Dataset (WESAD) [4].

The data consists of over **3.5 million** entries for each test subject. There are a total of 15 test subjects.

The signal data is divided into chest and wrist devices with each having Accelerometer (ACC), ECG (Electrocardiogram), EMG (Electromyography), EDA (electrodermal activity), Temperature, BVP (from photoplethysmograph (PPG)).

The data also included a label for the test subject's state of mind - baseline, stress, amusement, meditation.

Each of the sensors collected data at different frequencies. 5 of the 10 columns and the labels were recorded at 700Hz and hence the other values were synchronized by repeating them to get an equal number of rows for each sensor.

A. Factors/Variables

The entire series was plotted for some of the columns and some interesting trends/patterns were noticed.

ECG - The ECG graph was very periodic with peaks and troughs at regular intervals. Each of the regular peaks included

a small trough, a small crest, a large trough, a large peak finally followed by a small trough. This repeating pattern continued throughout the whole series. This can be seen in Figure 1.

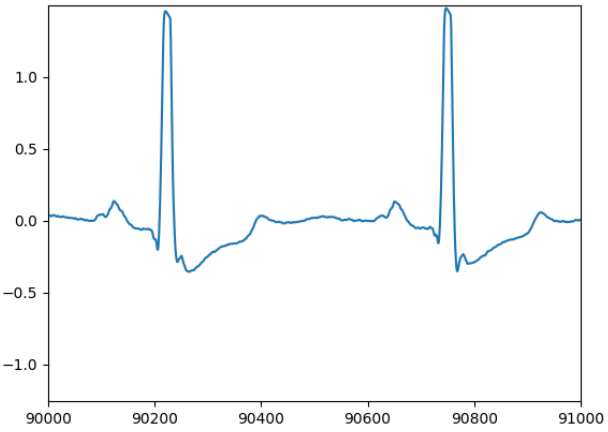

Fig. 1. A small part of the chest ECG series.

ACC - regular, well-defined changes i.e., the values change suddenly throughout the graph. ACC is the accelerometer readings which could mean that the sharp changes happen when the subject changes positions. For example, standing to sitting. Refer Figure 2.

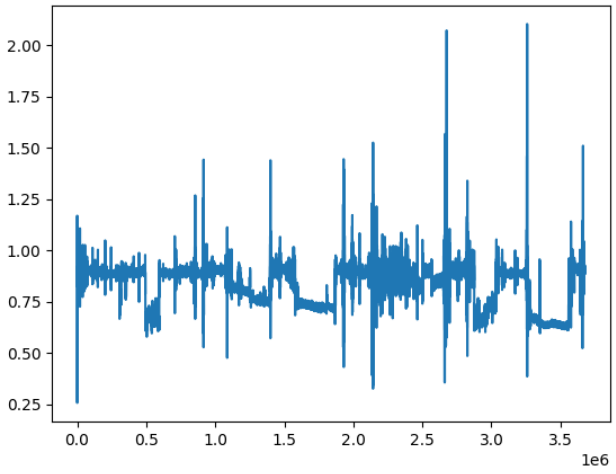

Fig. 2. The entire chest ACC series.

EMG - As seen in Figure 3, EMG shows regular trends/variations which correspond to the specific periods of recordings (”stres”, ”amusement”, ”meditation”). In particular, EMG values decrease considerably during meditation or amusement and increase when under stress.

Temp - regularly changing. Reached a peak during the “stress” period and steadily decreased. Refer Figure 4.

On plotting the correlation matrix (Figure 5), there was noticeable correlation in these values:

- Chest_temp and wrist_temp (high negative correlation)
- wrist_EDA and chest_EDA (high positive correlation)

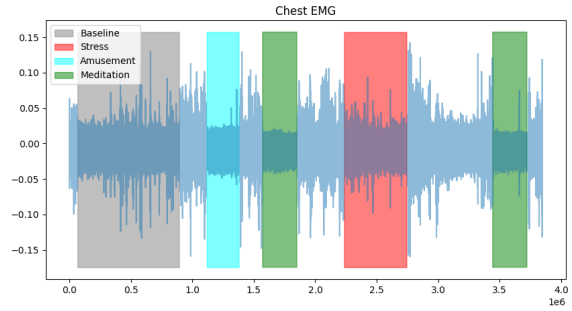

Fig. 3. The entire chest EMG series with interesting periods highlighted.

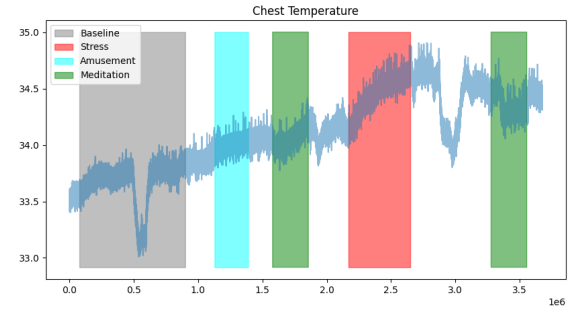

Fig. 4. The entire chest temperature series with interesting periods highlighted.

- Wrist_temp and chest_EDA (positive correlation)

There are no missing values in any of the columns and there is no inconsistent entries, this is likely due to the fact that the sensors were accurate all the time and did not malfunction. A few values at the beginning, after recording was started, behaved differently from the rest of the values. This might be due to the devices being installed on the test subjects while they were still recording.

The dataset was also checked for outliers (points outside *1.5*IQR*) in the ECG graph and found that over **8 lakh** of the points were outside that range which was due to the infrequent peaks in the ECG as discussed before. Refer Figure 6.

B. Comparison of factors across test subjects

It can be seen in Figure 7 that ECG has considerable variation across test subjects.

Figure 8 shows the variation of mean body temperature across different test subjects. Here again there is considerable variation.

Thus, we can conclude that each subject has different values for the given vitals.

IV. Principal Component Analysis

Principal Component Analysis, is a dimensionality-reduction method which is used to reduce the dimensionality (number of columns) of large data sets.

There were 10 features in the data set. It is known that to achieve clustering it is difficult to plot a dataset with 10

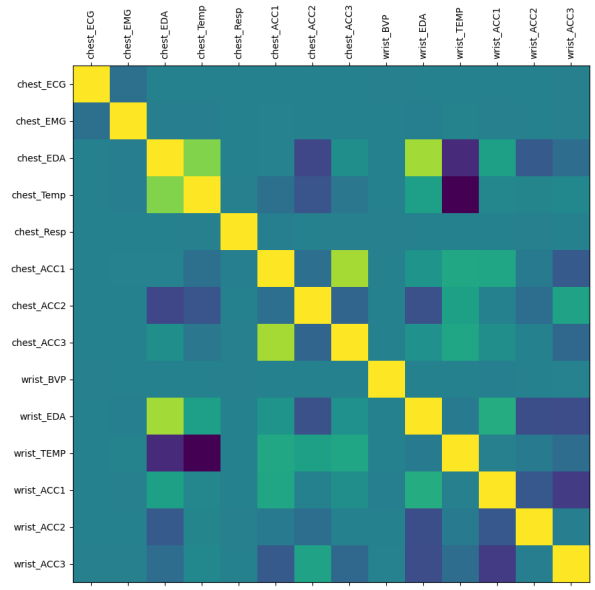

Fig. 5. The correlation matrix.

Fig. 6. Boxplot of chest ECG

features as there will be 10 dimensions.Therefore PCA was performed on the data set.

Initially we tried to perform PCA to reduce the dataset to 2 dimension and the final result as label but the PCA variance ratio that was achieved was not satisfactory as it can be seen from the plot.

It was decided to perform PCA for 3 dimension and the PCA variance ratio was above 95%. The dataset was reduced to 3 principal components which are Principal Component 1, Principal Component 2 and Principal Component 3. The label was concatenated with the final Data frame achieved. Rows which had a label of 1, 2, 3 or 4 were included in the PCA as these were the required labels for the further studies. (the numbers indicate: baseline - 1, stress - 2, amusement - 3, meditation - 4).

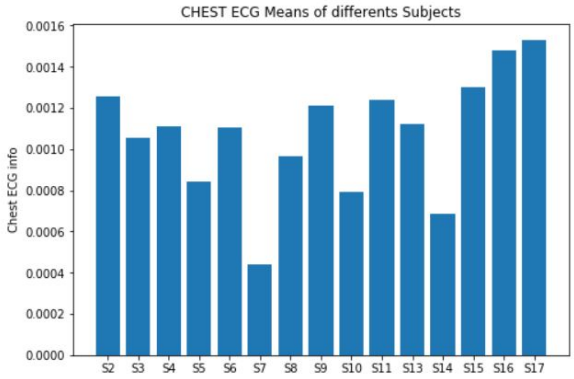

Fig. 7. Means of Chest ECG of different Subjects1

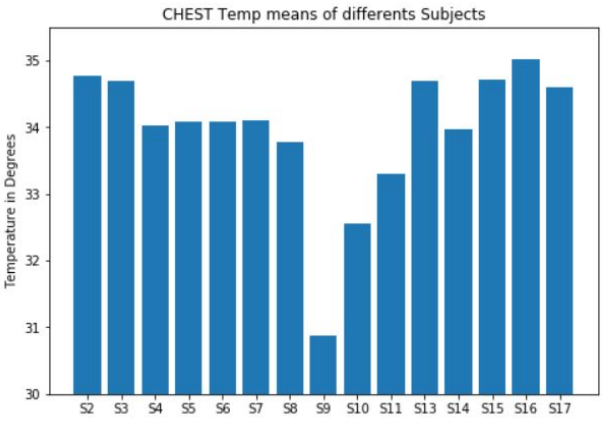

Fig. 8. Means of Chest Temp of different Subjects1

There are interesting observation obtained from the graph.

- It can be seen that a large amount of points are marked by purple colour which represents the subject is stressed most of the time during the study.
- The number of (mix of yellow and purple colour) points are less that implies that subject is not amused as such.
- A good amount of blue colour points are also noticed in the observation which suggests the mood of the subject in baseline.

Once the required number of Principal components are obtained, they are used for clustering, which is explained in the next section.

Figure 10 shows a visualisation of PCA in 3 dimensions for one test subject from a random sample of 10,000 points. We can infer from Figure 11, which shows the same for 100,000 points, that the random sample is representative of the dataset and can be used for modelling.

V. Clustering

A. K-Means

The first task is to find the number of clusters which are needed to classify the data. The elbow point technique is used to determine the number of clusters that can classify the

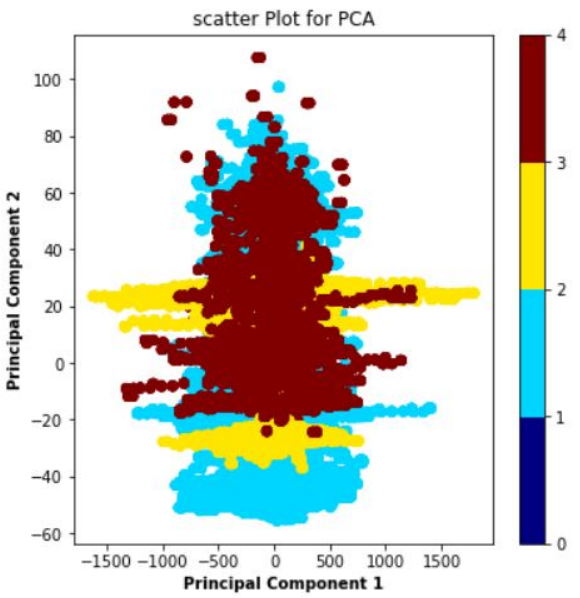

Fig. 9. Principal Component Analysis with 2 Components

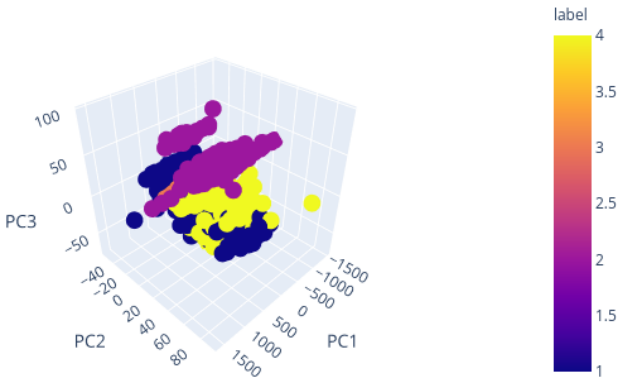

Fig. 10. Principal Component Analysis with 3 components with a random sample of 10,000 points.

dataset. From the figure 12 the elbow point lies on 3 which was expected also because after doing PCA on the dataset we know our data has three states for emotion i.e baseline, stress, amusement.

B. DBSCAN

Apart from K-Means, we also tried to use DBSCAN. Using the data generated after performing PCA, we ran DBSCAN on a sample of 10000 points to see how well it clusters. The sample data is plotted in Figure 14. After some parameter tuning, the value for *eps* and *minDistance* were *0.5* and *10* respectively. As shown in Figure 15 the generated 3 categories, with -1 being noise. The *yellow* and *orange* categories represent stress whereas *purple* category represents baseline. With this, we can infer that stress is relatively easier to differentiate from the other affects.

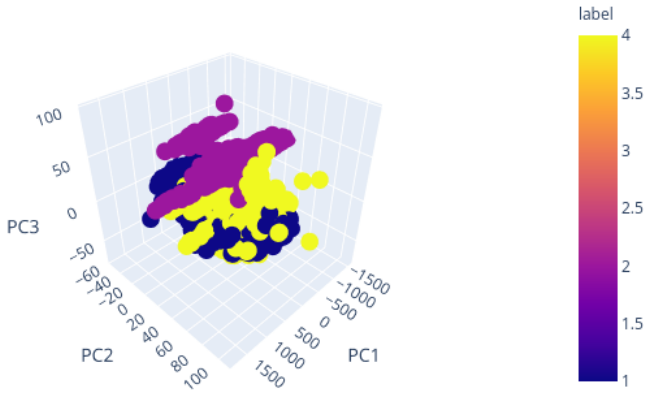

Fig. 11. Principal Component Analysis with 3 components with a random sample of 100,000 points.

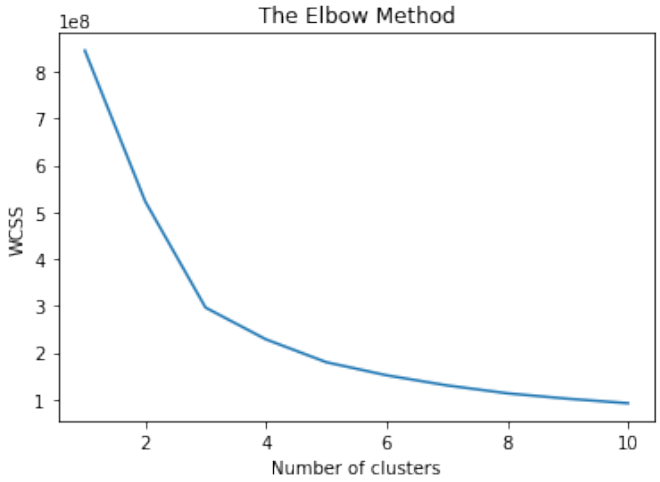

Fig. 12. The elbow point for 50,000 data points.

VI. Classification

Given that the dataset consists of labelled classes for each data point, it is evident that a supervised classification model can be trained to predict, given the sensor readings, the affect (state of mind) of the test subject.

As we have seen in Figure 10 and Figure 11, a random sample is representative of the entire dataset. Since there are 2 million valid data points for each subject and due to resource constraints, the classification algorithms were run on a random sample of 50,000 points.

Before using the classification models, we pre-process the data by first keeping only the points whose label is one of 1 (baseline), 2 (stress), 3 (amusement) and 4 (meditation). Then we randomly split the data into 70-30% of train and test data.

After training the classification models, we test them using the following metrics:

1) Train accuracy: accuracy on the train split
2) Test accuracy: accuracy on the test split
3) Precision: on the test split

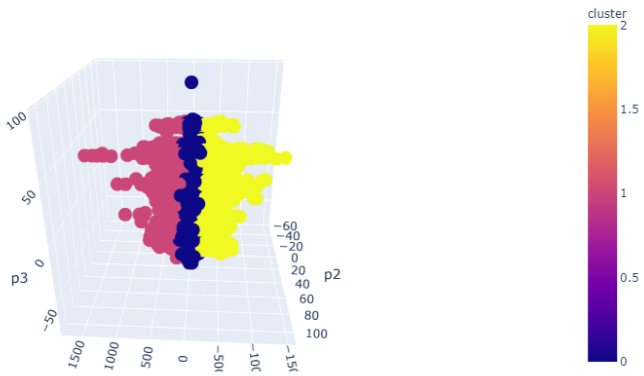

Fig. 13. Output of K-Means clustering

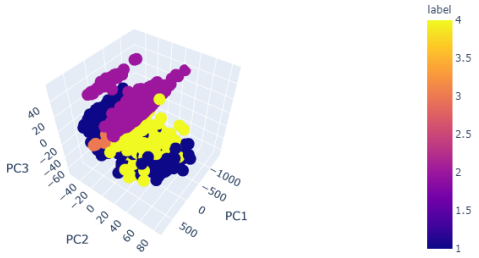

Fig. 14. Sample Data of 10000 points

4) Recall: on the test split
5) F1-Score: on the test split

We can see in Figure 16, Figure 17, Figure 18, Figure 19 and Figure 20 the metrics for various classification models tested.

It can be noted that all the models, except for AdaBoost, perform very well on the dataset, achieving over 95% accuracy as well as F1-score. SVM with linear kernel and random forest perform the best with near perfect test accuracy and F1-score.

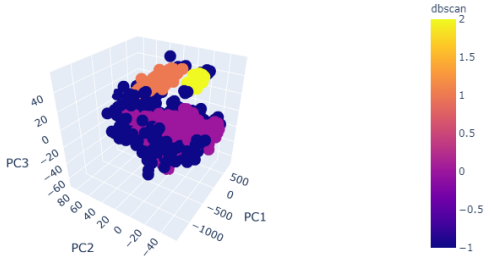

Fig. 15. Output of DBSCAN Clustering

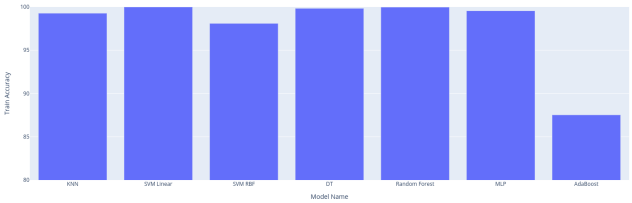

Fig. 16. Comparison of train accuracy of different classification models

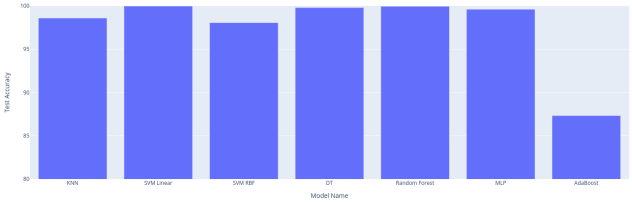

Fig. 17. Comparison of test accuracy of different classification models

It is also apparent that the train and test accuracies are nearly the same for all models (except for KNN) suggesting that there is no over-fitting and the dataset does not have much variation to make the test set much different from the train set.

Details of models tested:

1) K-Nearest Neighbours (KNN) classifier with $k = 3$
2) Support Vector Machine (SVM) classifier with linear kernel
3) Support Vector Machine (SVM) classifier with Radial Basis Function (RBF) kernel
4) Decision Tree (DT) classifier with maximum tree depth 5
5) Random Forest classifier with maximum tree depth 5

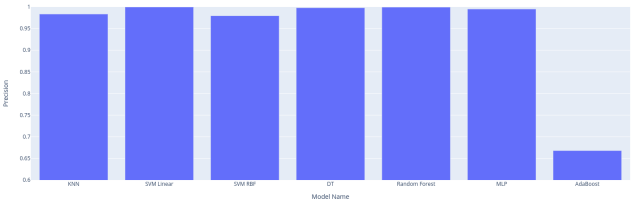

Fig. 18. Comparison of precision of different classification models

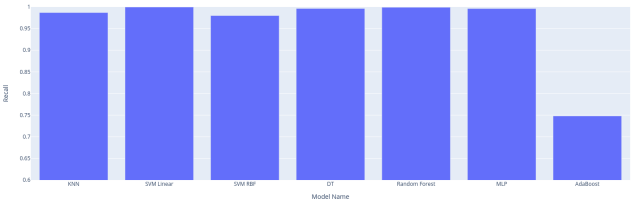

Fig. 19. Comparison of recall of different classification models

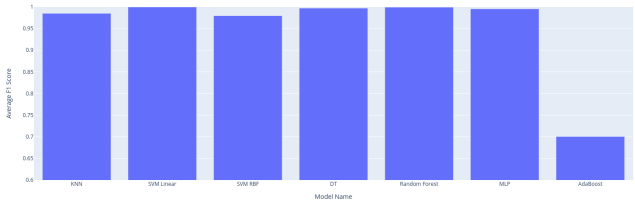

Fig. 20. Comparison of F1-score of different classification models

and 10 estimators

6) Multi-layer Perceptron (MLP) classifier with learning rate 1 and 1000 iterations
7) AdaBoost classifier

VII. Conclusion

With this paper, we provide a thorough exploration of the data with multiple visualisations along with comparison between classification models. The following conclusions are drawn from our work.

- Repeating patterns were found in ECG graphs, sharp changes were found in accelerometer graphs, regular trends corresponding to stress level were found in the EMG graphs and finally, peaks of temperature were noticed during stress periods.
- Using the Principal Component analysis, the number of components can be reduced to 3. We first tried to reduce it to 2 components but the results were not satisfactory. Finally 3 Principal components were chosen.
- Using the DBSCAN clustering algorithm, we could differentiate the periods of stress from other periods in an unsupervised manner.
- Standard classification algorithms such as K-Nearest Neighbours or Random Forest classifier can be used to almost perfectly (nearly 99% accuracy and F1-score) determine a person's state of mind or affect using just the sensor readings.

References

[1] Hogg, Michael A. and Abrams, Dominic (2007) Social cognition and attitudes. In: Martin, G. Neil and Carlson, Neil R. and Buskist, William, eds. Psychology. Third Edition. Pearson Education Limited, pp. 684-721. ISBN 978-0-273-71086-8.

[2] Schmidt P, Reiss A, Durichen R, Laerhoven KV. Wearable-Based Affect Recognition-A Review. Sensors (Basel). 2019 Sep.

[3] O. Mozos, V. Sandulescu, S. Andrews, D. Ellis, N. Bellotto, R. Dobrescu, and J. Ferrandez. 2017. Stress detection using wearable physiological and sociometric sensors. International Journal of Neural Systems 27, 02 (2017), 1650041.

[4] Schmidt, Philip and Reiss, Attila and Duerichen, Robert and Marberger, Claus and Van Laerhoven, Kristof. (2018). Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. 400-408. 10.1145/3242969.3242985.

[5] 2016. HSE on work related stress. http://www.hse.gov.uk/statistics/causdis/-ffstress/index.htm. (2016). Accessed: 2017-09-06

[6] M. Fernandez-Delgado, E. Cernadas, S. Barro, and D. Amorim. 2014. Do We Need Hundreds of Classifiers to Solve Real World Classification Problems? J. Mach. Learn. Res. 15, 1 (2014), 3133–3181.

[7] M. Soleymani, M. Pantic, and T. Pun. 2012b. Multimodal emotion recognition in response to videos. IEEE Transactions on Affective Computing 3, 2 (2012), 211–223

[8] Hussein Al Osman, Tiago H. Falk. 2017. Multimodal Affect Recognition: Current Approaches and Challenges. DOI: 10.5772/65683.

[9] S. D'mello and J. Kory. 2015. A Review and Meta-Analysis of Multimodal Affect Detection Systems.ACM Comput.Surv.47, 3, Article 43 (2015), 36 pages.

[10] Fitri Indra Indikawati and Sri Winiarti. 2020. Stress detection from multimodal wearable sensor data. IOP Conference Series: Materials Science and Engineering.