Lifestyle Disease Detection using Cyber Physical Systems

Submitted By :-
Kartik Nema (IIT2018156)
Bhupendra (IIT2018163)
Prakhar Srivastava (IIT2018172)
Ashish Patel (II2018175)
Shubham Soni (IIT2018177)

UNDER THE SUPERVISION OF
Dr. Sonali Agarwal
Associate Professor

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, ALLAHABAD

(A UNIVERSITY ESTABLISHED UNDER SEC.3 OF UGC ACT, 1956 VIDE NOTIFICATION NO. F.9-4/99-U.3 DATED 04.08.2000 OF THE GOVT. OF INDIA)
A CENTRE OF EXCELLENCE IN INFORMATION TECHNOLOGY ESTABLISHED BY GOVT.OF INDIA

November, 2020

CANDIDATE'S DECLARATION

We do hereby declare that the work presented in this study entitled "Lifestyle disease detection using Cyber Physical System", submitted towards the fulfillment of the course ID20BMPR5C, in Information Technology at Indian Institute of Information Technology, Allahabad, is an authentic record of our original work carried out under the guidance of Dr. Sonali Agarwal. This report was done in full compliance with the requirements and constraints of the prescribed curriculum.

Place: Allahabad

Date:

CERTIFICATE FROM SUPERVISOR

Date:

I do hereby recommend that the semester project work prepared under my/our supervision titled "Lifestyle disease detection using Cyber Physical System" be accepted in the partial fulfillment of the requirements of the mini project(ID20BMPR5C) in Semester V.

Date: **Guide's name & Signature**

Place: Allahabad **Dr Sonali Agarwal**

Associate Professor

Index

List of Figures

Abstract - In this paper, we discuss our foundations to develop a solution which could help to detect lifestyle diseases (i.e. diseases caused by the way a person lives or his/her habits, example stress, obesity etc.) in the modern times using cyber physical systems. We mention related work in this domain and what we aim to do.

Keywords - Cyber physical systems, Lifestyle diseases, Stress, Logistic Regression, Confusion matrix.

1. Introduction

1.1 *Problem Definition and Objectives*

Our project title is “Lifestyle Disease Detection using Cyber Physical Systems”. For the purpose of this project, we have considered stress as the target disease. We have tried to develop a mode, which takes into consideration several features like biological responses of the person like GSR (EDA), ECG. It also records the form responses of the subject under different circumstances (stressing phase, relaxing phase and enjoyment phase). Finally we train the machine learning model using this data (available from the WESAD dataset) and make a prediction.

1.2 *Lifestyle Diseases*

As the name suggests lifestyle diseases are caused due to the lifestyle led by a person, i.e. how they lead their lives. There are several reasons for the rapid growth of these diseases across the last several years, some of them are unhealthy eating (junk food), no exercise, unhealthy habits like smoking etc. These diseases include obesity, diabetes etc. By nature of spreading these diseases are non-communicable.

1.3 *Cyber Physical Systems*

A cyber physical system (CPS) is a computer system in which a mechanism is controlled or monitored by computer-based algorithms. In CPS, physical and software components are deeply intertwined, able to operate on different spatial and temporal scales, exhibit multiple and distinct behavioral modalities, and interact with each other in ways that change with context [11]. Examples of CPS include smart grid, autonomous automobile systems, medical monitoring, industrial control systems, robotics systems, and automatic pilot avionics.

Cyber-Physical Systems (CPS) are integrations of computation, networking, and physical processes. Embedded computers and networks monitor and control the physical processes, with feedback loops where physical processes affect computations and vice versa.CPS integrates the dynamics of the physical processes with those of the software and networking, providing abstractions and modeling, design, and analysis techniques for the integrated whole.

1.4 *Motivation*

In the ever increasing speed of the modern world, stress has become a more and more common disease. Stress levels in India are much higher than several of the developed nations, around 82% of Indians suffer from stress [3]. The most stressed age group is the 35 - 49 years old.

Extreme levels of stress can have disastrous effects on the health and well being of humans, it could result in higher blood pressure (hypertension), depression etc. Acute stress is the stress which easily goes away, however as the level of stress goes on increases it might transform into chronic stress, which is hard to rectify or even manage. Via our project we aim to find a simple method to detect levels of stress in an individual, thus being able to provide an early warning to them in case their stress levels happen to be high.

2. Literature Review

(i) Activity-Aware Mental Stress Detection Using Physiological Sensors[1]

(a) Purpose:-Researchers make up for the impacts of physical exercises by extracting a lot of accelerometer features that identify unique physical exercises alongside EEG and GSR features.

(b) Methodology:-Researchers chosen the 20 participant and ask them to do these 3 tasks-

1. Baseline segment (10 minutes): Listen to meditation music (in seated, stand- ing, or walking position).
2. Mental task segment (10 minutes): Complete Stroop test and mental arithmetic under time pressure while seated, standing, or walking.
3. Recovery segment (10 minutes): Sit in a chair with closed eyes and listen to meditation music.

Researchers recorded six data sets(and each segment contains 19200 GSR samples and 60000 ECG and accelerometer samplesSize of the whole date -set for 20 participants was 45 mb (because of continuous recording of data).

For each participant they recorded the 60 minutes of data and divided it into 1 minute intervals(windows).

Sensor	Features
ECG	Mean RR, Std RR, Mean HR, Std HR RMSSD, pNN50, LF, HF, LF/HF ratio
GSR	Mean SCL, Std SCL, Total magnitude, Duration, and Number of startle responses
Accelerometer	Mean of X, Y and Z axis Standard deviation of X, Y, and Z axis Energy of X, Y, and Z axis Correlation coefficient of XY, YZ, and ZX

TABLE-Features Used in Model

mean HR: mean heart rate (beats per minute);
mean RR: mean heartbeat interval (ms);
SDNN: standard deviation of RR-intervals between normal beats;
RMSSD: root mean square of the difference between successive RR-intervals
pNN50: the percentage of heartbeat intervals with a difference in successive heartbeat intervals greater than 50 ms
LF (0.04-0.15 Hz): a low-frequency component that is mediated by both the SNS and PNS;
HF (0.15-0.4Hz): a high-frequency component mediated by the PNS;
LF/HF: LF to HF ratio that is used as an index of autonomic balance
SCL-Skin conductance level

After that researchers normalize the dataset and researchers classified the data using various classification methods.
Since heart rate changes very rapidly between activities so the results were improved when they have not used ecg data. GSR was not affected much between activities and when researchers included the GSR accuracy was improved. Best classification accuracy (92.4%) was obtained by using the decision tree classifier with the all-features.

(c) Results:- Researchers found that the accelerometer improves the accuracy of the model. Researchers achieved 92.6% accuracy using all the features and GSR data was independent in all activities hence it's a better indicator.

(ii) Investigating heart rate variability: a machine learning approach [2][7]

The dataset is derived from the PHD thesis presented by Jennifer A. Healey [7]. The original data collection was done by testing on young people during stressing environments, like rush hours or driving on highways, several parameters were noted including ECG, EMG, HR, GSR etc. This paper serves as a binary classifier, i.e. whether the person is stressed or unstressed. The use of python automated machine learning libraries TPOT to create a model is discussed.

The model is tested with new data obtained from wearable devices like Fitbit.

To make prediction whether the person is stressed or not, the major parameters considered are :-

(a) Galvanic Skin Response sensor :- GSR is one of the most effective factors to measure stress. If stress is more than the person sweats more, i.e. conductance increases and resistance decreases. An obvious problem with using GSR as a sole perimeter is that people stress even due to climatic conditions, disregarding the stress levels.

(b) Heart rate variability :- HRV is a useful parameter for monitoring stress. HRV refers to the feature extracted from the RR interval (time interval between two consecutive heartbeats). During times of stress there are changes in heart rate (HR) and heart rate variability (HRV).

Furthermore the data was labeled to indicate stress levels, i.e. there was no feature to tell whether the person is actually stressed or not originally, for this purpose another feature was added, which will have value 1 to signify stress and 0 unstressed. This was done by measuring GSR of feet, for all the instances, then finding the median of the GSR value series and using it as a cutoff. Any instance with value greater than the cutoff would be labeled stressed.

For testing purposes the subject was made to watch a stress provoking tv show. This could be for example a horror show, or any scene that can trigger stress. The wearable devices are used to continuously monitor the subject during this period. The subject is asked to log all the scenes which caused stress to him/her. Finally they compared the log of the subject a and the output the classifier produced at that time.

(iii) Stress Detection Using Low Cost Heart Rate Sensors.[8]

(1) Purpose:- The paper presents the ideas and consequences of two investigations focused at stress identification with a low cost heart rate sensor, a chest belt. In the gadget approval study, we thought about pulse information and different highlights from the belt to those deliberate by a best quality level gadget to evaluate the dependability of the sensor.

(2) Methodology:- In this study two steps involved that are follows:-

(a) Device Validation Study:**-**We tried the dependability of an ease telemedical heart rate sensor against an acknowledged clinical gadget.

It further includes following steps:-

(i) Sensor Selection and Measurement Protocol:- Here in this paper, they have analyzed various devices that detect heart rate and finally chosen the device named CardioSport TP3 Heart Rate Transmitter device, it is a simple device that detect heart rate and millisecond accurately and it does not have its own memory due to which for storage purpose they have selected the Nexus 7 tablet with Android version 4.4.2 to connect to the device with the bluetooth 4.0 protocol and store the measured data on the tablet.

(ii) Synchronization Procedure:-They utilized a basic methodology to synchronize the information estimated by the CardioSport gadget with those deliberate by the highest quality level gadget so as to encourage their comparison. The calculation utilizes a sliding window that goes from the earliest starting point of the chest belt sign as far as possible and computes the supreme mistake between the two signals. When sliding is finished, the area of the sliding window with the base supreme blunder is considered as where the two signs ought to be synchronized.

(iii) Statistical analysis and data processing:- Here they performed time and frequency domain analysis and obtained the correlation and mean absolute error of the measurement.

(b) Clinical study:-We performed and assessed a clinical report, utilizing the approved telemedical sensor.

Steps involved are:-

(i) Measurement Protocol:-46 healthy volunteers, mostly university and high school students (27 men and 19 women; average age: 24.6 years), participated in the experiment. The experiment was divided into two parts with a duration of 10 minutes each, so the whole procedure lasted for 20 minutes.

In the initial segment, the members were approached to attempt to unwind in an upstanding sitting position while tuning in to unwinding music. The second aspect of the examination was a psychological errand intended to fill in as a wellspring of mental pressure. Then they were asked to play a game called Stroop color test.

(ii) Statistical analysis:-We examined and thought about information utilizing the MedCalc programming and Microsoft Excel.

(iii) Stress Detection Algorithm:-It uses time domain HRV features instead of frequency domain to lessen the computational power. We used a combination of the mean HR, pNN50, and RMSSD features to identify stress. We used a brute force approach to find the threshold for the HRV feature. This algorithm is unable to detect rest state. So for convenience the rest state is considered if stress was not detected. The flow chart of the algorithm is given in figure 1.

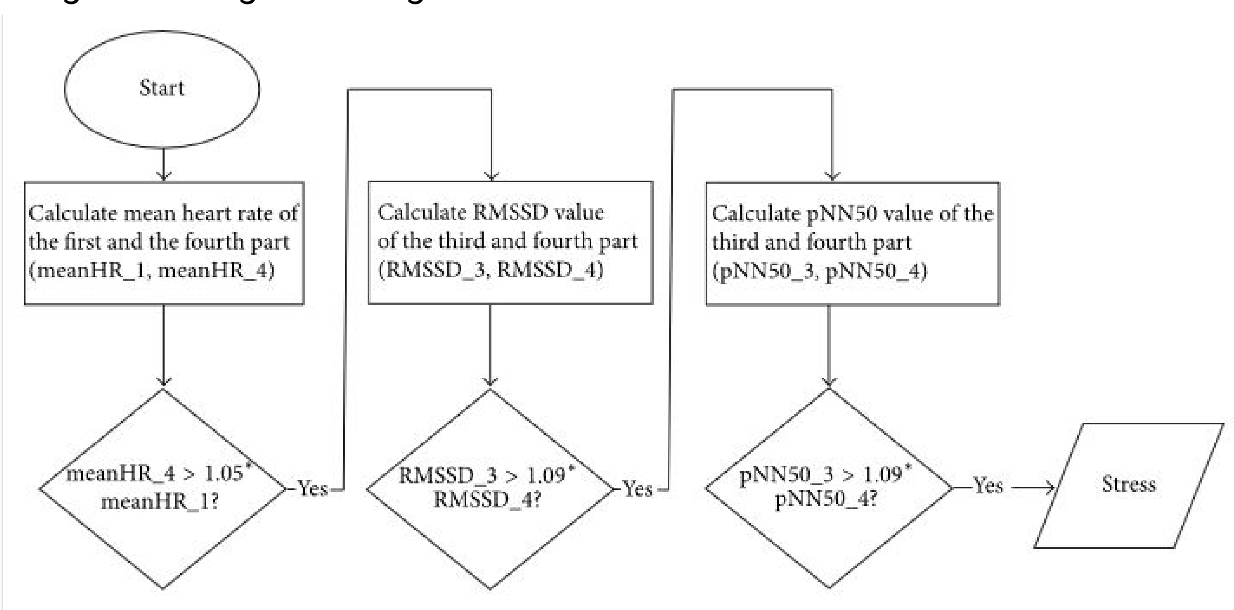

Figure 1. Flowchart of the method used in this paper [8]

(3) Results:-The accuracy, sensitivity, and specificity values for correctly detecting stress are 74.60%, 75% and 74.19%. A progression of distressing events rather than a solitary significant event can likewise continuously place the subject into an stressed condition. This model may conclude a false statement that only the last event has made the stress without considering the previous events.

3. Research Gap

The available methods use a test kit which measures the ECG,GSV and HRV values.These are recorded and then a machine learning algorithm is applied which predicts whether a subject is stressed or not.

This is an expensive and time taking process since each subject needs to wear the devices and then the data is recorded.To reduce the costs and save time we decided to implement a method in which each subject have two options:

1. Get an ECG/GSV/HRV test done which will give more accurate results.
2. Get stress score from a questionnaire which will give slightly less accurate results comparatively.

If a subject voluntarily chooses option 1 we measure the ECG/GSV/HRV values by showing him comedy videos and mental arithmetic videos.After that we could ask him to fill a questionnaire. This would be a simple stress detection questionnaire and it would have options. The responses given by the subject are recorded. Since we know the stress level of the subject we identify this as a label and features as the options.We train this model to give scores to each option.

If a subject opts for 2nd we just give him the questionnaire and we give him the stress score according to the responses he makes.

Since data collection was not possible therefore we just trained our model on the 70% of the WESAD dataset and tried to predict the stress level on the rest 30% of the dataset. We were able to achieve an accuracy of around 87.59% on the overall dataset and a mean accuracy of 83% on doing individually on each subject.

4. Associated Challenges

In the process to propose a methodology for this problem statement we faced some challenges. These included :-

(a) Selecting input parameters :- The common parameters used to predict stress in the above papers include GSR, HRV, ECG etc. We will now discuss these features one by one :-
(i) GSR (galvanic skin response) :- Refers to the change in the activity of sweat glands or the conductance of the skin in response to changes in our emotional state [4][2]. GSR is used to monitor the continuous changes in the activity of sweat glands or skin conductance. Many studies have shown there is a relationship between GSR and stress levels. The more the person is stressed more is the conductance, lower the resistance, greater sweat organs activity, greater value of GSR.
(ii) HRV (heart rate variability) :- It is a measure of variation in time between consecutive heartbeats. If a person is in a stressed situation or fight-or-flight mode, then the variation between successive heartbeats is low, i.e. HRV is low in panic situations. On the other hand in times of relaxation, HRV is high, indicating a healthy nervous system, a high HRV indicates balance. A low value of HRV is often indicative of stress, while the people with high HRV are more healthy. [5]

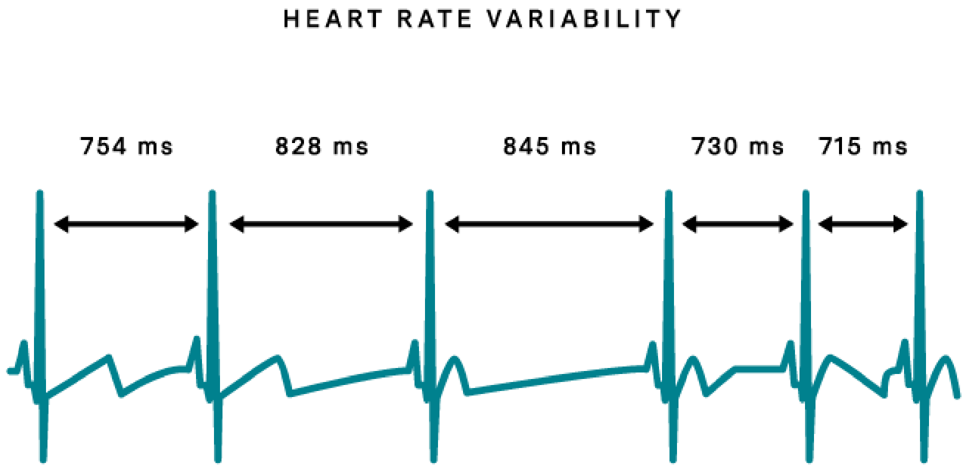

Figure 2. Heart Rate Variability [9]

(iii) ECG (electrocardiography):-It is used to measure the electrical activity of the heart. It is used to measure heart health. It is used to check heart issues like irregular heart rhythms.
ECG gives a graph as output where time is on x-axis and voltage(amplitude) on y-axis.usually reading are divided in squares where
1 big square consist of 25 mini squares
1 big square =0.5mV (on y-axis)
1 big square =0.20 seconds

1.when there is regular rhythm in QRS complexes(peak) then

$$heart\ rate = \frac{300}{No.\ of\ large\ squares\ between\ two\ qrs\ complexes}$$

2.when rhythm is irregular heart rate is given as

$$heart\ rate = 10 \times (number\ of\ R\ values(peak)\ in\ 6\ sec.)$$

Some other features were also mentioned in some other papers, we needed to decide which features to use in our work.
Galvanic skin response (GSR) is one of the most important features used for the purposes of stress detection. In research conducted by (María Viqueira Villarejo, Begoña García Zapirain, and Amaia Méndez Zorrilla) [2][5], where they achieved an accuracy of 76.56 % by using GSR as the sole parameter.

5. Implementation

The data is collected and then we use the 29 features which are BVP,EDA,Temperature,Respiratory and their minimum,maximum and mean values. Apart from these, age and weight are also taken as features.The label is taken as the stress level(0,1,2).

5.1 *Code description*

Our code implementation has two parts i.e one is for preprocessing the raw data and other one is for classification of the feature vector. The data is recorded in the raw form which needs to be preprocessed so that our classifier can be trained on the data. The preprocessing code(preprocessing.py) uses the data which is stored in the pickle files. The pickle files are available for each subject. The data from it is extracted and stored in m14_merged.csv. Then we have Mini.ipynb which is our classifier and uses the random forest classifier.

5.2 *Dataset description*

Our method uses a WESAD dataset which is a publicly available dataset for wearable stress and affect detection.The data is recorded from a chest(RespiBAN) and wrist(Empatica E4) worn device of 15 subjects. The sensors record blood volume pulse, electrocardiogram, electrodermal activity, electromyogram, respiration, body temperature, and three-axis acceleration. Empatica E4 records the blood volume pulse (BVP), body temperature, electrodermal activity(EDA),etc. RespiBAN records electrocardiogram(ECG),EDA,respiration,temperature and 3-axis acceleration.

5.3 *Technologies used*

Language :- Python (Python 3)

Tools and Libraries :-

For development purposes we used coding environments like Google Colab, PyCharm, Jupyter Notebook.

A number of text editors like sublime text, Atom, Notepad++, VS code etc. were used.

Libraries include sklearn, numpy, pandas, matplotlib, seaborn, os, pickle etc.

5.4 *Software & Hardware Requirements*

The code was developed (in addition to Google Colab) on local system having the following specifications :-

RAM :- 8 GB

Intel Core i5 processor

1 TB hard disk drive

Operating Systems :- Windows 10 (64 bit), Linux Ubuntu (LTS 18.04)

For interpreting and running the code on the local system a terminal or command prompt are required. In Windows 10 command prompt (cmd) and in Linux Ubuntu (bash terminal) were used. Anaconda, a cross platform tool could also be used for these operations, it also provides direct access to the other tools like Jupyter notebook.

Running the code (on local system) :-

To run the code python 3 was first installed in the system. The required libraries were then installed using pip (package installer). Softwares like pyCharm were installed as it makes maintaining the code much easier.

For collaboration tools like github and slack were used.

5.5 *Finished project description*

The project consists of 2 major parts

(i) Data Cleaning

1. **preprocessing.py**

The data for each subject is stored in pickle files.This code extracts the data and converts it to a simple csv file which will have all the features.

Step 1: A class is created which will have attributes as the properties of a subject(signal,label and subject),types of signals(chest and wrist),data recorded by chest device(ACC,EMG,EDA,Temp,etc.) and data recorded by wrist device(ACC,BVP,EDA,TEMP).

Step 2: This class has member functions which returns the wrist data and chest data and a function which creates a map which will have key as the type of data recorded in the wrist device and value as the variance of the data recorded.

Step 3: The data from the pickle files is loaded for each subject.

Step 4: For each data the minimum,maximum and standard for BVP,EDA(phasic),EDA(smna)(which is calculated using cvxEDA.py),Temperature is calculated.Also the age and weight of the subject are extracted from the pickle file.

Step 5: BVP peak frequency is calculated using the built in periodogram function in scipy.signal module.

Step 6: These data are then pushed into the dataframe and all subjects data is merged into one.

Step 7: The final dataframe is written into the m14_merged.csv. This file is the cleaned dataset which is used for classification.

2. **cvxEDA.py:**

This file is used to concatenate new features from EDA.This is a Convex Optimization Approach to Electrodermal Activity Processing[12].The approach defined in [12] gives the phasic,tonic and an additive white Gaussian noise term incorporating model prediction errors as well as measurement errors and artifacts [12].

(ii) Classification

The WESAD dataset cannot be used for classification in its original structure, thus we need to preprocess it. WESAD contains separate directories and sub directories for each and every subject, amounting to a total size of greater than 2 GB. Upon running the data cleaning code, we obtain the data in a ready to use format (csv).

For the purposes of classification of data, we first load in the data from the csv file and feed it to our machine learning model. For this purpose we use the Random Forest algorithm.

We have provided the csv file, so that the data can be readily available instead of downloading the WESAD dataset and then extracting the features any time somebody wants to run the code.

The code for classification is available both as python notebook (.ipynb) and python (.py) file. The python script can be run directly on the local system which has python3 installed, while the notebook can be run by uploading it on a platform like Google Colab or Jupter notebook.

6. Results

First we trained our model on the 70% of the WESAD dataset and tried to predict the stress level on the rest 30% of the dataset. We were able to achieve an accuracy of **94.0677966101695%** on the overall dataset upon using the Random Forest classification algorithm.

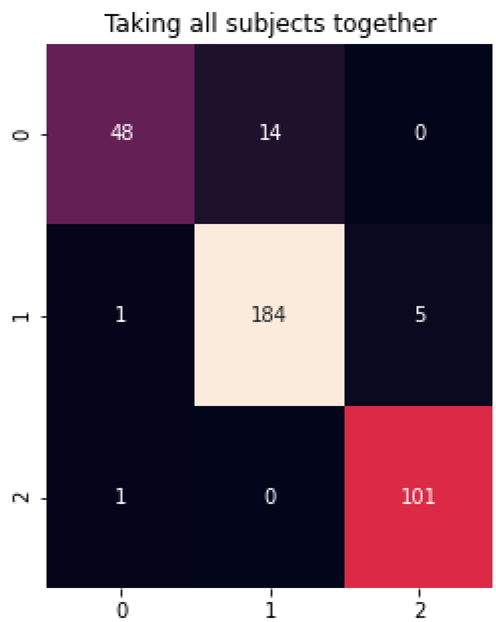

figure 3 : Confusion matrix generated upon testing on 30% of the dataset

Code Metrics:

1. Using Random Forest Classifier

   F1 Score: 0.9255834371813753
   Precision Score 0.9473743726573916
   Recall Score 0.9109368931500161

2. Using Decision Tree:

   F1 Score 0.7507779994557787
   Precision Score 0.7673336651733864
   Recall Score 0.7401000921024891

Accuracy	F1 Score	Precision Score	Recall score
94.067796610165%	0.925583437181375	0.9473743726573916	0.9109368931500161

Next we tested our model on all the 15 subjects individually. The accuracy for each subject was computed, and finally the mean accuracy was calculated. In our dataset the subject IDs range from 2 to 11 and 13 to 17. The data regarding subject id 1 and 12 was removed from by the owner of this dataset, due to some ambiguities.

The mean accuracy obtained was 98.33236714975845 %

For testing the model upon every subject, we take the data corresponding to each subject id and split it into training and testing parts in the ratio 7:3. We now test our model using this testing, and obtain the following results.

Subject ID	Accuracy
2	100.0 %
3	91.66666666666666 %
4	95.65217391304348 %
5	100.0 %
6	100.0%
7	91.66666666666666 %
8	100.0 %
9	100.0 %
10	96.0 %
11	100.0 %
13	100.0 %
14	100.0 %
15	100.0 %
16	100.0 %
17	100.0 %

Table depicting the accuracies obtained for predictions corresponding to all the 15 subjects

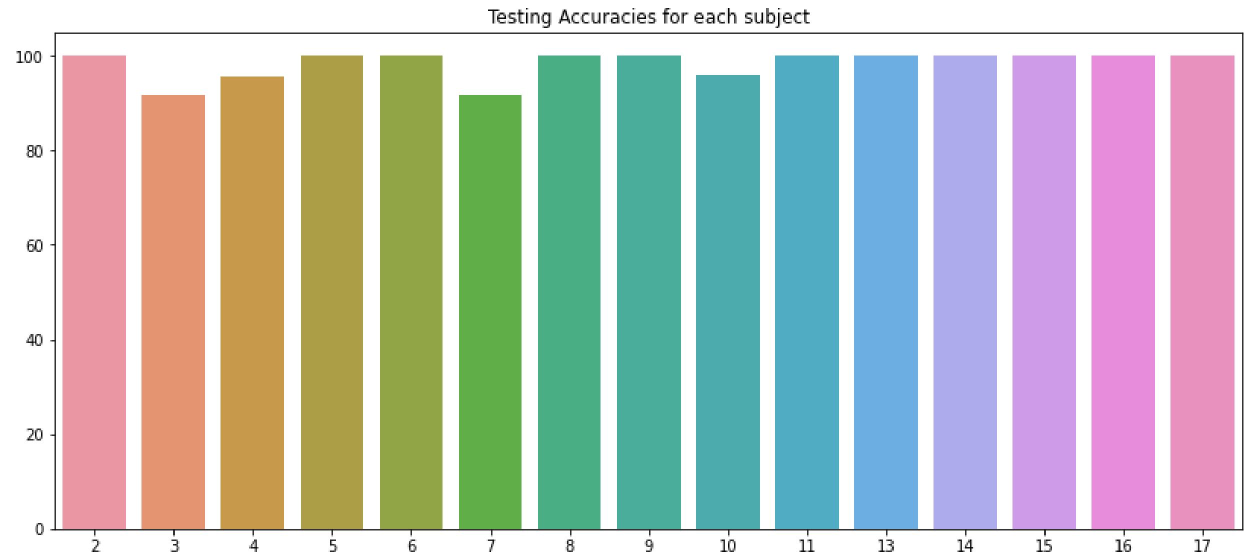

Figure4 : Histogram depicting the accuracies for subjects

Figure 5 : Confusion matrices for all the subjects

7. Activity Schedule

(i) **C1 Duration:-**

S.no.	Activity	Duration
1.	Project Title Deciding	21 Aug - 26 Aug (6 days)
2.	Analyzing Papers for Reference	27 Aug - 4 Sep (9 days)
3.	Identifying an Overall Structure for how the Project should look like and Functions	5 Sep - 11 Sep (7 days)
4.	Concluding Final Methodology	12 Sep (1 day)
5.	Making Report and Presentation	13 Sep - 15 Sep(3 days)

(ii) **After C1 Duration:-**

In this duration we have done the implementation part and model testing for different datasets.

S.no	Activity	Duration
1.	Training the Model using the Data Collected	25 Sep - 10 Oct(15 days)
2.	Testing the Model Using the Score Given by ECG/GSR/HRV Method	11 Oct- 20 Oct(10 days)
3.	Using the Model Created to Calculate Stress Scores	21 Oct - 31 Oct(11 days)
4.	Testing Model for Different DataSets	1 Nov- 9 Nov(9 Days)
5.	Final Report and Presentation	10 Nov- 15 Nov (6 Days

8. Conclusion

The proposed method uses a stress detecting questionnaire and gives score to each option of the questions by asking some volunteers to take an ECG/GSR/HRV test to detect the stress level and making them attempt the questionnaire. After getting scores for each option our questionnaire is now ready for being used to calculate the stress level. Every candidate wanting to have a stress test is asked to opt for a full stress test using ECG/GSR/HSV or have a free questionnaire test. If the candidate opts for 2nd option he is given the questionnaire and stress level is calculated using the score for each option.

Since we do not require everyone to have the ECG/GSR/HSV test this method is portable and cost effective but accuracy might decrease if the candidate does not attempt the quiz optimally.

9. References

[1] Feng-Tso Sun1 , Cynthia Kuo1,2, Heng-Tze Cheng1 , Senaka Buthpitiya1 , Patricia Collins1 , and Martin Griss1, “Activity-Aware Mental Stress Detection Using Physiological Sensors”, Carnegie Mellon University, https://eudl.eu/pdf/10.1007/978-3-642-29336-8_16

[2] Christopher André Ottesen, “Investigating heart rate variability: a machine learning approach”, School of Electronic Engineering and Computer Science, University of London.
https://onedrive.live.com/?authkey=%21AHxruWzJ4fI3JJM&cid=C2AB9C8D5071989A&id=C2AB9C8D5071989A%21112359&parId=C2AB9C8D5071989A%21112358&o=OneUp

[3] Stress related data :-
https://economictimes.indiatimes.com/wealth/personal-finance-news/82-indians-bogged-down-by-stress-cigna-360-well-being-study/articleshow/68615097.cms

[4] GSR :-
https://imotions.com/blog/gsr/

[5] María Viqueira Villarejo, Begoña García Zapirain, and Amaia Méndez Zorrilla, “A Stress Sensor Based on Galvanic Skin Response (GSR) Controlled by ZigBee”
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3386730/

[6] https://imotions.com/blog/heart-rate-variability/

[7] Jennifer A. Healey, “Wearable and Automotive Systems for Affect Recognition from Physiology”, Department of Electrical Engineering and Computer Science MASSACHUSETTS INSTITUTE OF TECHNOLOGY
https://dspace.mit.edu/handle/1721.1/9067

[8] Mario Salai, István Vassányi, and István Kósa, “Stress Detection Using Low Cost Heart Rate Sensors”, Medical Informatics R&D Centre, University of Pannonia, Egyetem Utca 10, Veszprem 8200, Hungary https://www.hindawi.com/journals/jhe/2016/5136705/#results

[9] Heart Rate Variability, https://ouraring.com/what-is-heart-rate-variability

[10] MySignals Kit, https://www.libelium.com/wp-content/uploads/2016/09/mysignals_sensors_big.jpg

[11] https://en.wikipedia.org/wiki/Cyber-physical_system

[12] https://ieeexplore.ieee.org/document/7229284
