National Technical University of Athens

Interdisciplinary Master's Programme in

Data Science and Machine Learning

Biodata Analysis

Detection and Classification of Depression Using Data from Actigraph Biosensors

Professor

Konstantina Zarkogianni

Students

Αλευρά Ελένη AM 03400163

Θεοφίλη Σταυρούλα AM03400166

Κοτσιφάκου Κοντιλένια AM 03400174

Παπανικολάου Μαρία AM 03400188

1 Abstract

Depression is a severe mental disorder with characteristic symptoms like sadness, the feeling of emptiness, anxiety and sleep disturbance, as well as general loss of initiative and interest in activities. One common symptom of depression is altered motor activity, which can manifest as either increased or decreased movement. Actigraph recordings can be used as an objective method for assessing motor activity and its patterns, which can provide insights into certain aspects of depression. Using actigraphs, machine learning researchers can cooperate with healthcare professionals so as to gain objective information about motor activity patterns in individuals that indicate depression. For example, decreased motor activity and reduced movement during the day, known as hypoactivity, can be indicative of depression. In this project, we try to develop 4 different models in order to classify and categorize depression based on measurements of actigraph watches, demographics and phychological tabular data. Two of these models focus on predicting if someone is depressed or not, and the remaining two categorize the depressed patients as bipolar or unipolar depressive. The explainability of the models utilizing actigraphy and/or psychometric data is done only in the case of 70% or more accuracy in conjunction with relative explainable measures and techniques. Furthermore, there is a brief research on the field of explainable multimodal neural networks and a discussion about the results of the explainability techniques used.

2 Introduction

Depression has been a worldwide concern for a long time and continues to plague the global health agenda. According to the World Health Organization (WHO), more than 350 million individuals are estimated to suffer from depression which is equivalent to 4.4% of the world's population. Dealing with this mental health disorder can be very demanding since it can create physical, economical and emotional problems often leading to problems with work and sick leaves. However, early diagnosis followed by appropriate treatment has proven to be successful in reducing its impacts. Therefore, methods and tools for monitoring mental health are an immediate requirement.

The use of on body sensors to monitor personal health has become quite normal nowadays. The aforementioned sensors can be worn on the body, embedded in clothing or accessories, or even integrated into implantable devices. They typically utilize different types of sensors to measure and record data such as heart rate, blood pressure, body temperature, activity levels, sleep patterns, and more. As a result of their use, modern people can collect vast amounts of data every day, for purposes such as increasing quality of life, supervise their fitness levels, or even to change bad habits.

In the spirit of contributing to the later, not only by ameliorating one's self-care but also by creating models that can be used by an not-AI expert doctor to diagnose such a disorder. In this project, we created 4 explainable models which will be able to detect if a person is depressed or not, as well as the kind of depression that they suffer from. In order to do this, we used data collected from measurements of actigraph watches. An actigraph watch measures activity by using a piezoelectric accelerometer that is programmed to record the integration of intensity, amount and duration of movement in all directions. The sampling frequency is 32Hz, only movements over 0.05 g are recorded and the number of counts is proportional to the intensity of the movement.

To begin with, we developed two 1D convolutional neural networks (CNN) in order to extract some features for sequential data. Afterwards, we used all of these features so that it is able for us to train two different feed forward networks which predict if someone is depressed or not. The second one will also use as input the extracted features, as well as age,gender, education, admission to hospital, marriage, work status and some handcrafted data (min, max, mean, standard deviation). Furthermore, the previous model aims to inform us about the type of a patient's depression based on all these multimodal features. Finally we proceed to the development of two XGBoost models. One of them classifies depressed and not depressed people, based on mean, max, min, sd, and the second one will categorize the kind of depression receiving as input both handcrafted and tabular data. Before proceeding to model creation, in each case, we will perform data balancing in order to achieve the best model performance possible. Last but not least, using 10-fold cross-validation we compute and compare the accuracy of each corresponding model.

3 Related Work

Psychological analysis can be carried out using various methods. In this section, we discuss about some of the previous works performed using various techniques in order to detect if a person is depressed or not.

To begin with, *facial markers* are extensively considered in depression diagnosis due to the following reasons: first, depressed individuals tend to have anomalous facial manifestations for e.g., fewer smiles, more frequent lip presses, prolonged activity on the corrugator muscle, sad/negative/neutral expression occurrence, fast/slow eye blinks, etc. Second, several tools are now available to extract visual features, like The Computer Expression Recognition Toolbox[1], OPENFACE[2] etc. Back in 2018, *Wang et al*.[3] have examined the facial cue changes between depressed and normal subjects in the same situation (while displaying positive, neutral, and negative pictures). To measure the facial cue changes on the face, they used a person-specific active appearance model[4] to detect 68 point landmarks. Statistical features are extracted from distances between feature

points of eyes, eyebrows, corners of the mouth to feed the SVM classifier. The classifier achieved 78% test accuracy but the authors didn't focus on the explainability of their model.

Acoustic features of speech also play a vital role in diagnosis of depression for the following reasons: First, linguistic features (what subject speaks), paralinguistic features (how subject speaks), etc., are generally affected by the subject's mental state. Second, the clinician uses verbal indicators. Several studies have found distinguishable prosodic features such as pitch, loudness, energy, formants, jitter, shimmer, etc., between depressed and non-depressed individuals. *Cummins et al*.[5] have investigated good discriminative acoustic features that distinguish normal and depressed speakers. Features like Spectral centroid frequencies and amplitudes were computed using Mel-frequency Cepstral Coefficients (MFCC) then normalized. Multidimensional feature sets, i.e., combinations of those features have performed better when compared to single-dimensional features. They employed Gaussian mixture, but not explainable models to predict depressed and normal speakers. Further, Cummins[6] analyzed the effects of depression manifesting as a reduction in the spread of phonetic events in acoustic space. Him and his team found that depressed groups often tend to have reduced vowel space when compared with healthy people.

A quite similar approach is using *linguistics* in depression detection as it shows that words used by non depressed and depressed people may differ. In 2014, *Nguyen*[7] studied two online discussion groups, namely, "control" and "clinical" groups. The "control" group comprised people with a similar interest and fun-loving people, whereas the "clinical" group comprised people suffering from mental illnesses such bipolar disorder, major depressive disorder, SAD, and anxiety attacks. People in the clinical group discussed their issues freely and took advice for medication and intervention. The author found a difference in online communities involving people of these two groups. People in the "clinical" group usually use first-person pronouns ("I", "me," and "my") in comparison with "control" group people, who use fewer first-person pronouns and discuss various activities such as dancing, singing, and running.

Finally, some researchers have tried to combine *different modalities* due to the following reasons: first, an individual modality's contribution can be better understood when the convergence of modalities is carried out. Second, each modality has its own advantages. Hence a combination can yield a better outcome. For example, *Alghowinem* et al.[8] showed that the fusion of different modalities gives an improvement when compared to the individual modalities at hand. Their aim was to develop a classification-oriented approach, where features were selected from head pose, eye gaze, and verbal indicators of the depressed and healthy groups. The classification of these feature sets achieved the best results through the use of the SVM classifier. Though, these researchers, as the previous ones, did not use methods in order to extract the explainability of their model.

While the previous papers added knowledge to literature, they lack of explainability methods. As a part of this assignment it's important to elaborate more on techniques dealing with explainability. To start with, our first research involves studying and understanding *Joshi et al*.[9] paper. The later work includes thorough research and discussion on recent multimodal methods used for deep learning. The authors stated that there is a rising quest for model explainability, more so in the complex tasks involving multimodal AI methods. Their paper extensively reviews the present literature to introduce a comprehensive survey and commentary on the explainability in multimodal deep neural networks, from several topics on multimodal AI and its applications for generic domains, including a respectful amount of medical research. After organizing the explainable methods for deep learning models based on level and the scope of explainability, on feature attribution, distillation and intrinsic methods, there is a section devoted to multi-modal explainable techniques. It is important to provide more details of these approaches as a part of our project.

In the previous paper, it is referred that scientists used a variety of interesting attention-based techniques. Attention mechanisms played a crucial role in the alignment and fusion across different modalities. However, the authors implies that attention's explanatory power is questionable as they lack an association with the attention weights and gradients mapping for generating faithful explanations because attention does not explain if the model attends the right area. On the other hand, multi-modal explanations based on counterfactuality provide recommendations that result in actionable insights, recourse and a way to determine biases. Additionally, there are some graph based approaches utilizing scene and knowledge graphs. The scene graph for an image is the graphical representation of its contents where the nodes are the depicted objects, and the edges are the relationships between them. These are used so as to generate model explanations. Moreover, knowledge graphs significantly increase explainability through semantic information and domain knowledge base infusion. Last but not least there are some attribute-based solutions. Authors stated the significance of attributes in providing explanations. There are efforts in this direction, improving user trust and model flaws by generating complementary multimodal explanations for deep learning models using attributes and attribute maps.

While almost every one of the latter techniques, except attention, seems to be appropriate for our project, we decided to focus more on counterfactuals in respect to our familiarity to the subject (it was one of the proposed methods that was discussed in the lectures) and wide use of the method in other medical research. As a matter of fact, one of these papers is *Guarrasi at al.*[10] where authors present a deep architecture, explainable by design, which jointly learns modality reconstructions and sample classifications using tabular and imaging data. The explanation of the decision taken is computed by applying a latent shift that simulates a counterfactual prediction revealing the features of each modality that

contribute the most to the decision and a quantitative score indicating the modality importance. Their approach was tested on AIforCOVID dataset, which contains multimodal data ,X-ray and clinical data, for the early identification of patients at risk of severe outcome. The results of the paper illustrated that this method can provide meaningful explanations without degrading the classification performance.

4 Dataset and Features

The dataset contains actigraphy data collected from 23 unipolar and bipolar depressed patients (condition group) as well as actigraphy data from 32 non-depressed contributors (control group), consisting of 23 hospital employees, 5 students and 4 former patients without current psychiatric symptoms. There are actually two folders, whereas one contains the data for the *controls* and one for the *condition* group. For each patient, sensor data over several days of continuous measuring and some demographic data are provided.

The columns of each patient's csv contain: *timestamp* (one minute intervals), *date* (date of measurement) and *activity* (activity measurement from the actigraph watch). At this point, it is important to mention that the kind of activity that an actigraph watch measures is physical activity. It captures information about the intensity, duration, and frequency of bodily movements, including walking, running, jumping, and other forms of exercise. The accelerometer within the watch detects changes in acceleration and translates them into activity counts or steps, providing an objective measure of the wearer's movement levels. It can be observed that the condition group presents less activity level, especially during weekends.

In addition, the dataset contains another file named "scores.csv" which includes all the MADRS scores. The MADRS score (Montgomery-Åsberg Depression Rating Scale) is a standardized assessment tool commonly used in clinical and research settings to measure the severity of depressive symptoms in individuals diagnosed with major depressive disorder. It consists of a series of 10 items that assess various aspects of depressive symptoms, such as sadness, inner tension, sleep disturbances, appetite changes, concentration difficulties, and suicidal thoughts. Each item is rated on a scale from 0 to 6, with higher scores indicating more severe symptoms. The total score on the MADRS can range from 0 to 60, with higher scores indicating greater severity of depression. Here's a general guideline for interpreting MADRS scores:

- 0 to 6: No depression
- 7 to 19: Mild depression
- 20 to 34: Moderate depression
- 35 or higher: Severe depression

The columns of the *"scores.csv"* are:

1. number (patient ID)

   Each patientID consists of the group that the patient belongs to (condition or control), as well as a particular number, different for each patient of the same group.

2. days (number of days of measurements)
3. gender (1 or 2 for female or male)
4. age (age in age groups: 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69)
5. afftype (1: bipolar II, 2: unipolar depressive, 3: bipolar I)

   Bipolar disorder and unipolar depressive disorder (also known as major depressive disorder) are two distinct mental health conditions that involve mood disturbances. People with bipolar disorder type 1 experience episodes of mania that last for at least seven days or are severe enough to require immediate hospitalization. On the other hand, bipolar disorder type 2 involves recurrent episodes of major depression and hypomania. Hypomania is a less severe form of mania characterized by a distinct period of elevated mood and increased energy but with milder symptoms than full-blown mania. The hypomanic episodes in this type do not typically cause significant impairment or require hospitalization. Finally, unipolar depressive disorder involves persistent feelings of sadness, loss of interest or pleasure, changes in appetite and sleep, fatigue, difficulty concentrating, feelings of worthlessness or guilt, and in severe cases, suicidal thoughts. Unlike bipolar disorder, it does not involve episodes of mania or hypomania.

6. melanch (1: melancholia, 2: no melancholia)

   Melancholia is a subtype of major depressive disorder (MDD) characterized by severe depressive symptoms such as depressed mood, anhedonia (the diminished ability to experience pleasure or find enjoyment in activities that were previously pleasurable), anorexia and weight loss, disturbed sleep patterns.

7. inpatient (1: inpatient - a person who is admitted to a hospital or healthcare facility for an extended period, 2: outpatient - a person who does not require admission to a hospital or overnight stay)

8. edu (education grouped in years),

9. marriage (1: married or cohabiting, 2: single)
10. work (1: working or studying, 2: unemployed/sick leave)
11. madrs1 (MADRS score when measurement started),
12. madrs2 (MADRS score when measurement stopped).

5 Methods

To start with, there were 55 patients in the provided dataset with some corresponding days of actigraph data. In order to augment our dataset, there was an effort towards the use of artificial data, by adding a sensible amount of noise in the existing data. Nonetheless, that method ended up failing, with models not converging, resulting in not adequate results. Some other way was needed to be found in order to enlarge the dataset, so as to be meaningful to use neural networks. After taking into consideration all the available options, we tried to extract data from each day of a patient as a potential new patient, with the same demographic and psychological profile. Due to the fact that we didn't take into account that test and training patients (the initial ones) shouldn't be mixed together, we had to restart the whole process from the beginning and experiment again from scratch.

In order to detect if a patient is depressed or not, as well as the type of their depression we proceed to the development of five different models. To start with, a 1D convolutional neural network (CNN) was necessary in order to automatically extract some features for activity measures. Feature extraction is a common technique used to increase the size and diversity of a dataset by applying various transformations to the existing data samples. While this technique is most commonly associated with image data, it can be applied to other types too, including one-dimensional (1D) data such as time series, signal data or serial measurements like the ones we have from the actigraph watches. Actually, we have motor activity measurements for every single minute, but the input to the 1D CNN we created is the average of 30 minutes for the period of one day. The aforementioned derives from the fact that our data was needed to be denoised and half an hour average smoothing technique is considered suitable. The extracted features of this model, is part of the input of the next model we develop.

Unlike traditional methods, 1D CNNs automatically learn hierarchical representations from raw data, eliminating the need for manual feature extraction. The Encoder-Decoder architecture, a variant of the CNN model, further enhances the feature extraction process. The Encoder component learns to encode the input data into a compressed representation, while the Decoder component reconstructs the original input from the encoded

representation. This architecture enables the network to learn compressed representations of the input data, facilitating better feature extraction and subsequent analysis.

After the extraction of features, we proceed to the development of a neural network which receives as input all of the extra features of the previous step. The goal of this model is to classify each person as depressed or not depressed. In order to achieve an even more accurate classification, we create a second model, which also predicts if someone belongs to the ‘condition’ or the ‘control’ group. The difference between the last two neural networks is that the second one uses as inputs the extracted features of the 1D CNN and all the handcrafted features which are: Mean, Sd, Max and Min. The handcrafted feature *mean* refers to the average value of the motor activity measurements we have. It provides an indication of the central tendency of the data and it was calculated by summing all the values and dividing by the number of data points. *Standard Deviation* (SD) measures the dispersion or spread of the data points around the mean. It quantifies the amount of variability in the data and it is calculated by taking the square root of the variance. *Max* represents the highest value in our data and it provides information about the upper limit or peak value in the data. Finally, *min* represents the lowest value in measurements. It provides information about the lower limit or bottom value in the data. Our initial goal was to observe the differences of the two models, compare them and determine whether it is useful or not to utilize only the extracted features themselves. It was vital to examine, if and how handcraft features can enhance the prediction of the models.

To continue, we develop our last two models using XGBoost. XGBoost, short for "Extreme Gradient Boosting," is a popular machine learning algorithm used for both regression and classification tasks. It belongs to the gradient boosting family of algorithms and is known for its speed, scalability, and high performance. It can effectively handle both binary classification (two classes) and multiclass classification (more than two classes) problems, so it’s a quite promising tool that will help us detect and categorize someone’s depression. Taking advantage of XGBoost’s ability to handle complex interactions, handle missing values, and provide accurate predictions we create our first model which receives as input only the calculated features: Mean, Max, Min, Sd. Its purpose is to make a prediction about a patient, having as output the answer to the question if they are suffering from depression or not.

Finally, our last XGBoost model is developed in order to categorize a patient’s depression. In order to achieve that, we use as input some of our tabular data (gender, age, being admitted to hospital or not, years of education, marriage and work status). In this way, we have created a model which predicts the category of depression that someone suffers from bipolar or unipolar depression. The analysis of those two classes has been conducted in the previous section. Using as inputs not only the four handcrafted features, as the previous

model did, but also the tabular data we described before. At this point, it's important to mention that this algorithm is designed to handle tabular data, where the input features are structured in rows and columns, so it is ideal for our case. The output is the prediction of a patient's type of depression.

Before closing this section, we should mention that in every case, before training each model, we conduct data balancing. Balancing the data is usually necessary when we have imbalanced classes in our classification problem. In this particular case, in our effort to classify depression, we observe that unipolar depressive disorder seems to appear much more often than the other class. The latter situation can lead to poor performance, as the model we develop may struggle to learn patterns and make accurate predictions for the minority type. Thus, we need to perform data balancing by equalizing the number of instances in each class so that we can improve the model's performance.

Finally, we wanted to compute the accuracy of each model using 10-fold cross-validation. This technique allows us to make the most efficient use of our data. With 10-fold cross-validation, each instance in the dataset is used for both training and validation exactly once. This helps maximize the information extracted from the available data and provides a more representative evaluation. The average accuracy obtained from this method provides a quite more reliable estimate of the model's performance than a single train-test split. However, we weren't able to implement this technique in all cases because of the limitation of hanging to split training and testing original patients and never let our model see patients from the test set.

6 Experiments/Results

In order to reduce the dimensionality of our numerical data, we used a simple 1D CNN with one convolutional layer. After creating an Encoder-Decoder, we tested the accuracy of the model for different numbers of features in the output a. We present the performance of our model after 2000 epochs of training, with a learning rate equal to 1-4 and the weight decay constant equal to 1-5. Graphs show the reduction of loss in relation with epochs (Fig 1-4) and in Table 1 the loss is collected.

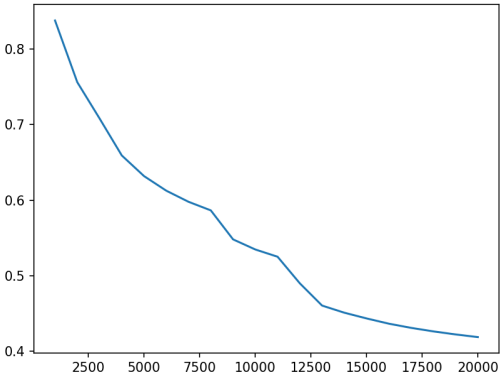

Figure 1. Loss reduction graph for a=15

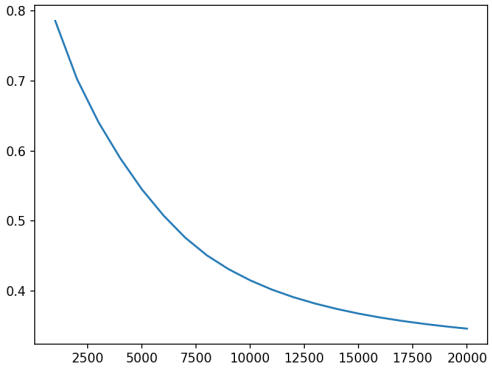

Figure 2. Loss reduction graph for a=20

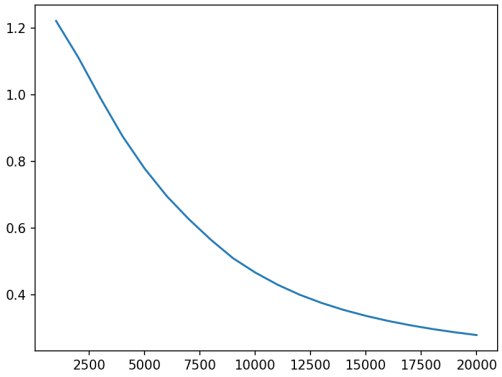

Figure 3. Loss reduction graph for a=25

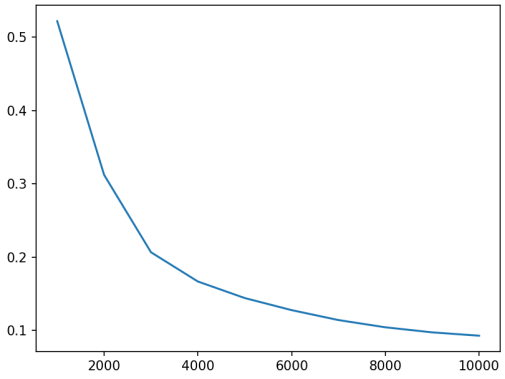

Figure 4. Loss reduction graph for a=30

a	**loss**
15	0.419
20	0.359
25	0.28
30	0.099

Table 1. Loss reduction table

In Fig. 5 the loss of the 1D CNN seems to decrease as the training epochs are increased.

```
Training model model1...
epoch: 1000 training loss:      0.522 validation loss:      0.440
epoch: 2000 training loss:      0.312 validation loss:      0.254
epoch: 3000 training loss:      0.207 validation loss:      0.176
epoch: 4000 training loss:      0.167 validation loss:      0.146
epoch: 5000 training loss:      0.144 validation loss:      0.132
epoch: 6000 training loss:      0.128 validation loss:      0.119
epoch: 7000 training loss:      0.114 validation loss:      0.105
epoch: 8000 training loss:      0.104 validation loss:      0.094
epoch: 9000 training loss:      0.097 validation loss:      0.087
epoch: 10000 training loss:      0.093 validation loss:      0.082
loss on the test set = 0.099
```

Figure 5. Training loss decrease

After optimizing the 1D CNN model to represent the initial features to 30 extracted ones, we repeat the procedure for the dataset where the classes distinguish the type of depression and the loss of best a is depicted in Fig. 6. The value of it is 0.177.

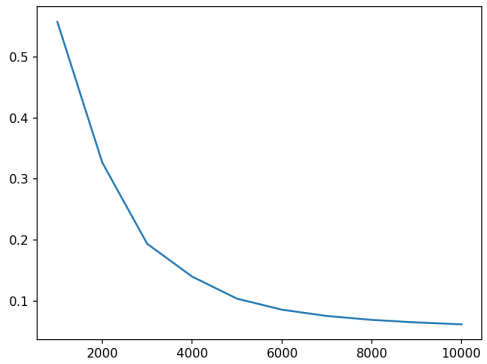

Figure 6. Loss reduction graph for a=30

As for the neural network model for detection of depression, we used multiple combinations of hyperparameters and we quote a table with their accuracy in relation to them. One example, the graph of the cost function decay as a function of learning iterations, as well as output, for epochs = 60000, lr = 0.01, weight_decay = 1e-05,momentum = 0.5 and lr_decay = 0.9 is illustrated in Fig 7.

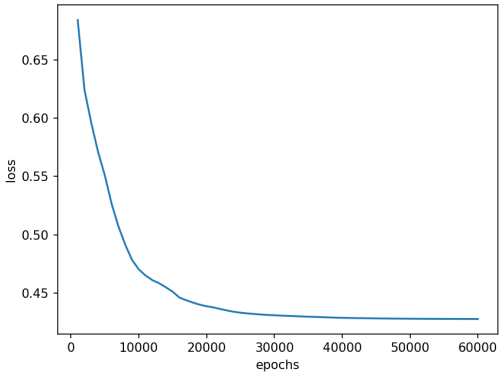

Figure 7. Loss reduction graph of neural network detecting depression with a specific set of hyperparameters (epochs = 60000, lr = 0.01, weight_decay = 1e-05,momentum = 0.5 and lr_decay = 0.9)

Unfortunately, despite our thorough investigation of hyperparameters, we didn't manage to have adequate accuracy (Figure 8).

```
loss on the original test set = 0.935
Accuracy on the original test set: 59.7000%

loss of the noisy test set = 1.158
Accuracy on the noisy test set: 57.4600%

The model is not sensitive to small changes in the input
```

Figure 8. Accuracy of the depression detection neural network model

The output of the model still wasn't good enough but this time having a bigger training set we are more confident about the reliability of the model to achieve the same performance after small changes of the input features. The results for the best hyperparameters are shown. in Fig 9. The scores for all the combinations of hyperparameters have been saved at outputs/model_1/NN_hyperparameters.csv.

```
lr: 0.0001, wd: 1e-05, mm: 0.9, lr_decay: 0.9, epochs: 10000
Epoch: 1000/10000, Train Loss: 0.6940, Validation Loss: 0.6678
-------------- Validation Accuracy: 0.7857, Test Accuracy: 0.5000
Epoch: 2000/10000, Train Loss: 0.6910, Validation Loss: 0.6578
-------------- Validation Accuracy: 0.7857, Test Accuracy: 0.5000
Epoch: 3000/10000, Train Loss: 0.6882, Validation Loss: 0.6510
-------------- Validation Accuracy: 0.7857, Test Accuracy: 0.5000
Epoch: 4000/10000, Train Loss: 0.6855, Validation Loss: 0.6454
-------------- Validation Accuracy: 0.7857, Test Accuracy: 0.5366
Epoch: 5000/10000, Train Loss: 0.6828, Validation Loss: 0.6401
-------------- Validation Accuracy: 0.7857, Test Accuracy: 0.5122
Epoch: 6000/10000, Train Loss: 0.6799, Validation Loss: 0.6349
-------------- Validation Accuracy: 0.7857, Test Accuracy: 0.5244
Epoch: 7000/10000, Train Loss: 0.6769, Validation Loss: 0.6294
-------------- Validation Accuracy: 0.7143, Test Accuracy: 0.5244
Epoch: 8000/10000, Train Loss: 0.6737, Validation Loss: 0.6238
-------------- Validation Accuracy: 0.7143, Test Accuracy: 0.5610
Epoch: 9000/10000, Train Loss: 0.6703, Validation Loss: 0.6181
-------------- Validation Accuracy: 0.7143, Test Accuracy: 0.5854
Epoch: 10000/10000, Train Loss: 0.6667, Validation Loss: 0.6122
-------------- Validation Accuracy: 0.6429, Test Accuracy: 0.5732
----------------------------
```

Figure 9. Accuracy of the best hyperparameters of the depression detection neural network model

Since, the accuracy is less than 70% and we didn’t manage to extract the explainability of the model. As it was pointed out in the lectures, it is not efficient to leverage the explainability of a non accuracy model.

Unfortunately, the same output was met after an equally extensive hyperparameter search for the detection type of depression model. Although our previous work had implemented counterfactuals for addressing the issue of multimodal explainability, in respect with the

proposed methods in literature, it wouldn't be beneficial to let doctors or other medical specialists rely on such models. Only for demonstration purposes, we attach one screenshot (Fig. 10) of counterfactual results without the previous analysis.

Diverse Counterfactual set (new outcome: 0.0)

	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	...	f29	f30	gender	inpatient	age	mean	std	max	min	target
0	0.230737	-3.535150e-11	0.617340	0.135508	3.017268e-11	-1.811280e-10	-0.001	0.378796	9.106870e-09	-6.759055e-09	...	-5.896983e-10	0.207510	1.0	1.0	1.0	0.435374	1.0	1.0	1.0	0
1	0.164868	-3.535150e-11	0.762884	0.306157	3.017268e-11	-1.811280e-10	-0.001	0.483216	7.037346e-02	-6.759055e-09	...	-5.896983e-10	0.216372	1.0	1.0	1.0	0.994747	1.0	1.0	1.0	0
2	0.386192	-3.535150e-11	0.892289	0.155406	3.017268e-11	-1.811280e-10	-0.001	0.348003	9.106870e-09	-6.759055e-09	...	-5.896983e-10	0.218201	1.0	1.0	1.0	1.000000	1.0	1.0	1.0	0
3	0.255818	-3.535150e-11	0.764296	0.164220	3.017268e-11	-1.811280e-10	-0.001	0.378449	9.106870e-09	-6.759055e-09	...	-5.896983e-10	0.428304	1.0	1.0	1.0	1.000000	1.0	1.0	1.0	0

4 rows × 38 columns

Warning: Total number of columns (38) exceeds max_columns (20) limiting to first (20) columns.

Figure 10. Depiction of invalid counterfactual results

While both neural network models didn't aim to precisely diversify one class over the other, that was not the case with XGBoost models. Following the steps of extensive hyperparameter grid search, the optimal hyperparameters were found for both of the two models. The case of depression classification, the best hyperparameters are included in Fig. 11.

```
Best hyperparameters:
{'subsample': 1.0, 'min_child_weight': 2, 'max_depth': 2, 'learning_rate': 0.05, 'gamma': 0.3, 'colsample_bytree': 1.0}
```

Figure 11. Best hyperparameters for the first XGBoost model

Even though XGBoost is a Machine Learning approach, the results were very promising and let us continue on the explainability part of the assignment, in contrast to neural network models. In fact the test accuracy of this particular model is 78.57%. Shapley values is the most suitable explainability method in this case.

Due to the fact that counterfactuals should not be analyzed, we are going to elaborate on the analysis of Shapley values and a potential use case of the model. In Fig. 12, 13 and 14 the exact numeric influence of each feature and the final predicted label is observed in three different ways. As a matter of fact, prediction of the input was successful, because the patient was classified as depressed. Experts leveraging an application that produces that kind of diagrams, can gain one's trust in the model, as they can evaluate if the decision of the model is based on rational factors or it is biased. Checking these factors can be done in any form depicted in the image, and the application may even personalize the diagram formulation based on the user's likings.

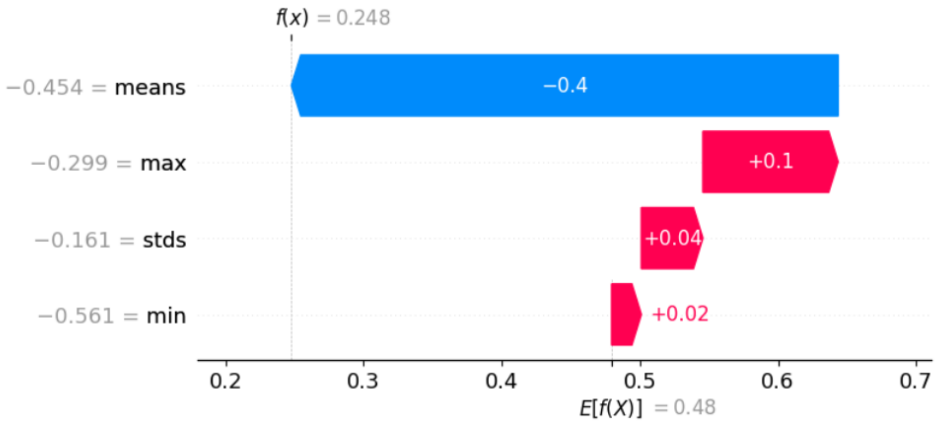

Figure 12. The effect of each feature in prediction (part a)

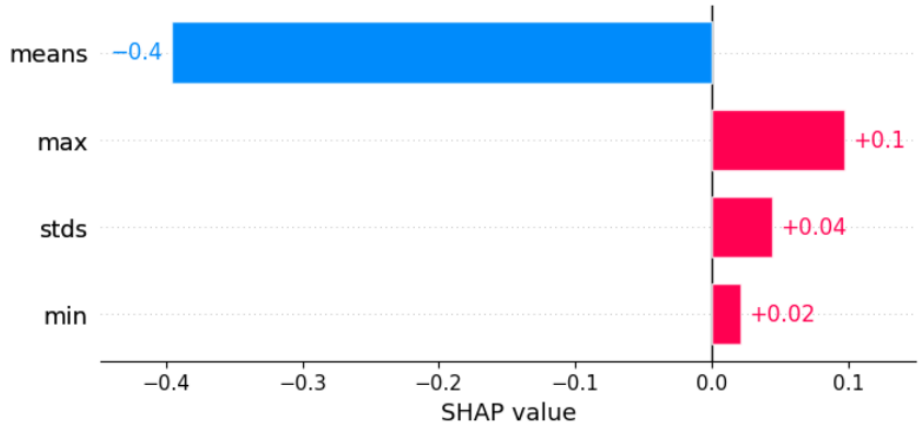

Figure 13. The effect of each feature in prediction (part b)

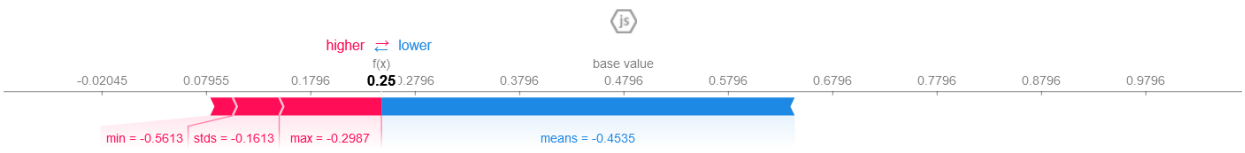

Figure 14. The effect of each feature in prediction (part c)

Suffice it to say that the XGBoost model attributes the depression state to the value of the mean of actigraph data. This result adds up with the fact that depressed people usually have less movement that not depressed people during the day. More sophisticated and established conclusions could be done by the experts utilizing the particular model we designed. In order to help the user paint the whole picture of model predictions, the live diagram illustrated in Fig. 15 can be used. Actually, it is written in JavaScript and it allows the user to move the mouse on a browser and observe the previously mentioned information for every patient.

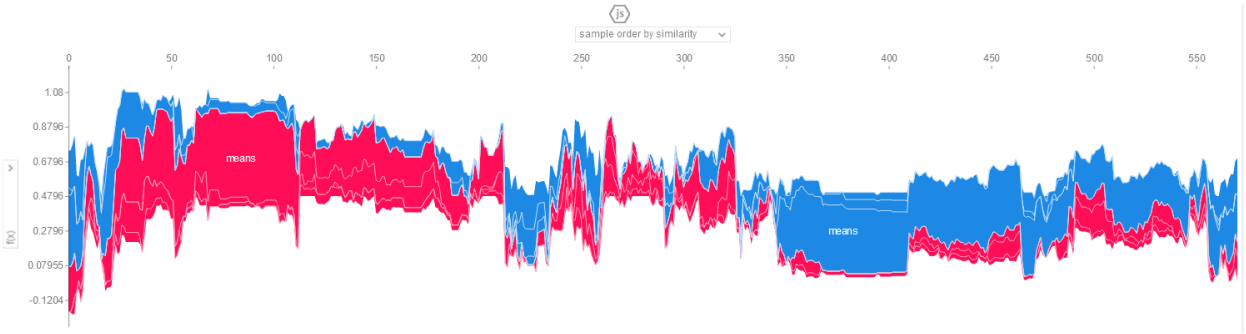

Figure 15. The effect of each feature in prediction for all the dataset

According to Fig. 16 the feature that plays a crucial role in labeling is the mean value of movement collected from actigraph watches. Moreover, the least contributing feature of the input seems to be the minimum value of the activity. It's not a fallacy to believe that the model can be trusted since almost all patients are going to have zero activity at some point of the day, perhaps in their sleep.

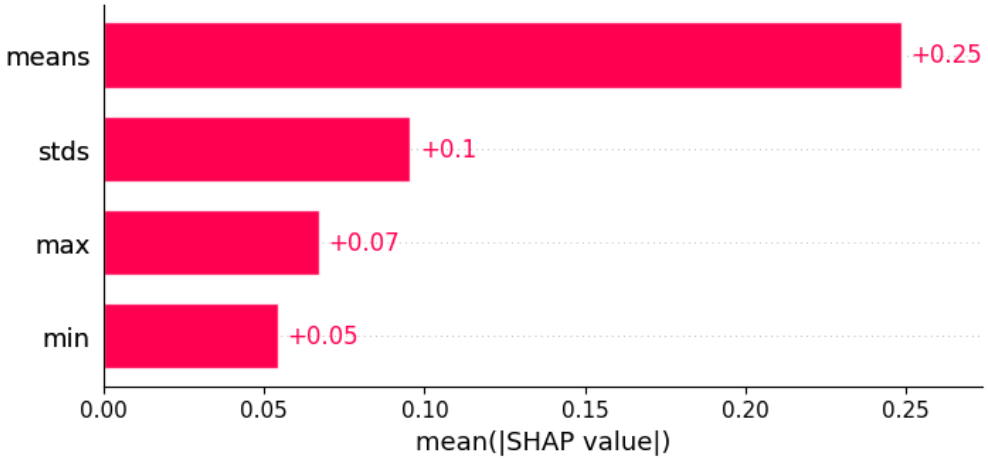

Figure 16. The absolute value of the effect of each feature in prediction for all the dataset

Additionally, mental health professionals can also extract useful information from the diagram in Fig. 17. Max value is quite obfuscated in the figure, and it seems to be difficult for a person to understand the logical explanation behind the prediction of the model. Lower values of mean lead to lower class label which is zero (depressed). While not everything is well explained, one expert could identify patterns and understand how the model works. In that case it could help him/her to trust or dismiss the model or even partially agree with it.

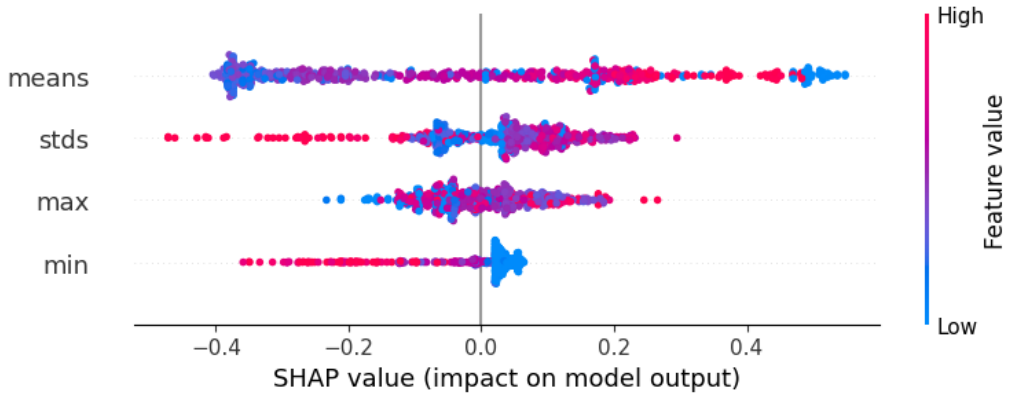

Figure 17. The influence of the volume of the value of each feature in prediction for all the dataset

Identical approaches were followed and similar results were found in the second XGBoost model, where the accuracy was even bigger (84.65%) and the classification problem of depression types was addressed. The best hyperparameters are shown in Fig.18.

```
Best hyperparameters:
{'subsample': 1.0, 'min_child_weight': 2, 'max_depth': 2, 'learning_rate': 0.05, 'gamma': 0.3, 'colsample_bytree': 1.0}
```

Figure 18. The influence of the volume of the value of each feature in prediction for all the dataset

The following example depicts the three forms of the effect of each feature in the case of a bipolar patient (class 0) in Fig. 19, 20 and 21. Features in the first case are sorted by the volume of the effect of labeling. For the particular patient, it seems that the fact that he/she is admitted to hospital, their age, and gender were the three more strong indicators of the class of the model. Work status mean and max have zero saying in the result. An expert can identify if there are biases or the model is close to accurate.

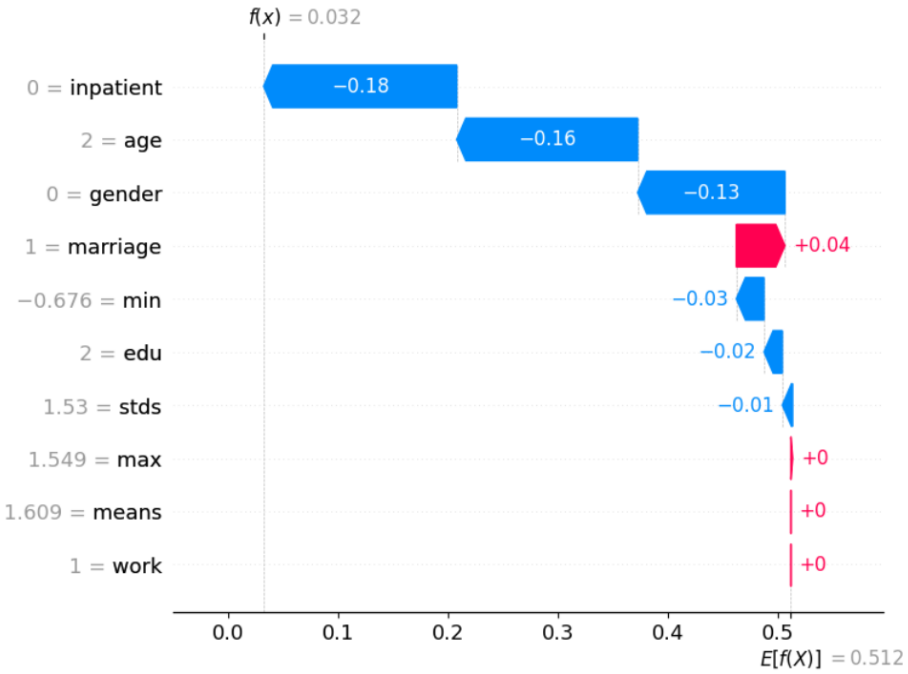

Figure 19. The effect of each feature in second model prediction (part a)

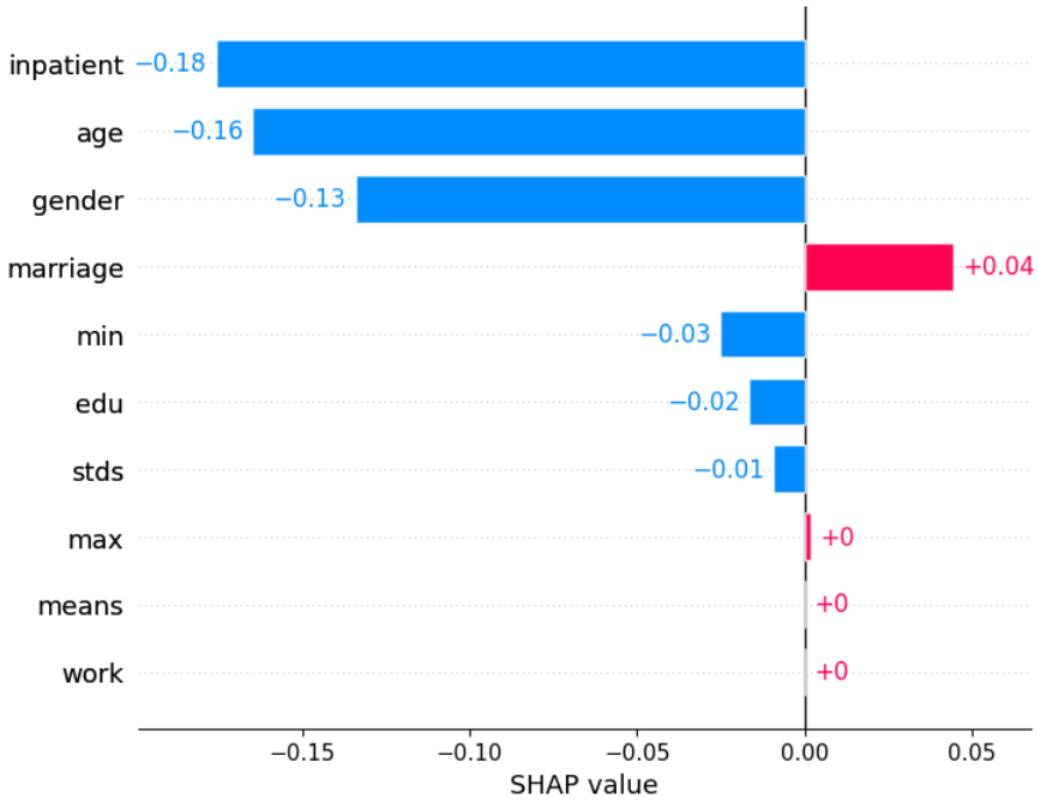

Figure 21. The effect of each feature in second model prediction (part b)

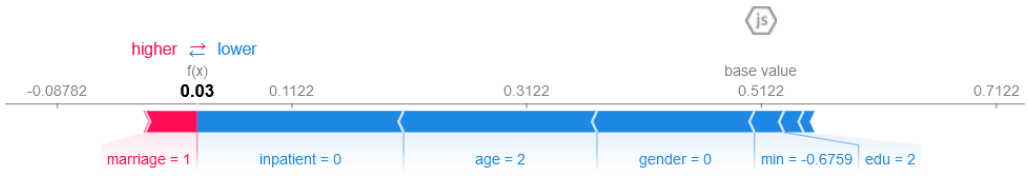

Figure 22. The effect of each feature in second model prediction (part c)

Once again, users can observe the overall influence of the features in the entire dataset (Fig. 23) or they can try to identify individual patients (Fig. 24). The results indicate that the age factor is at least twice as important for the type of depression based on the model. While mean value was thought as important in the previous model in order to identify a person with depression, it seems like it has no effect on the type of it, nor does the work status.

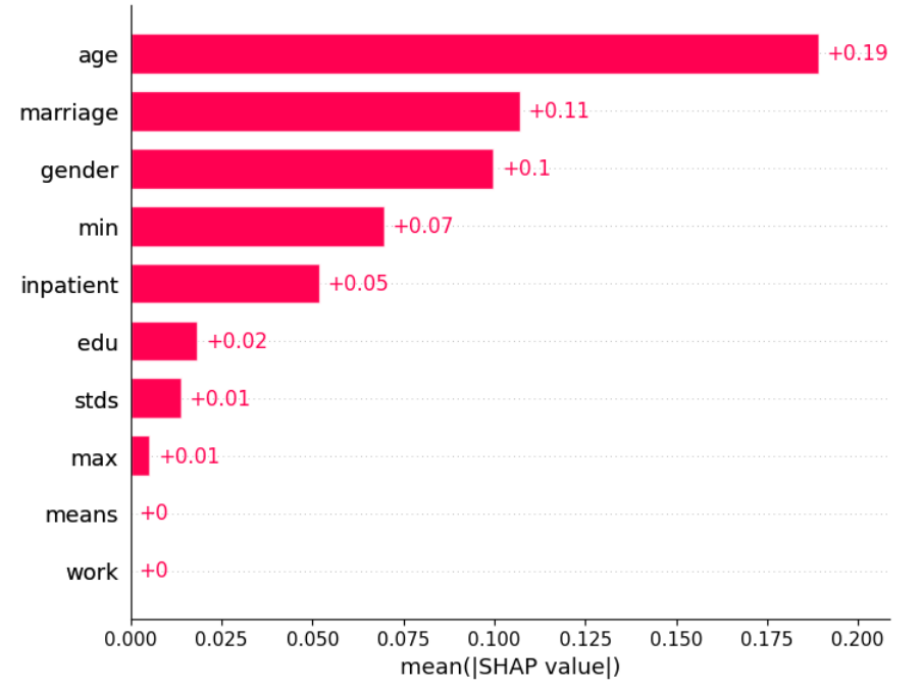

Figure 23. The absolute effect of each feature in the overall second model prediction

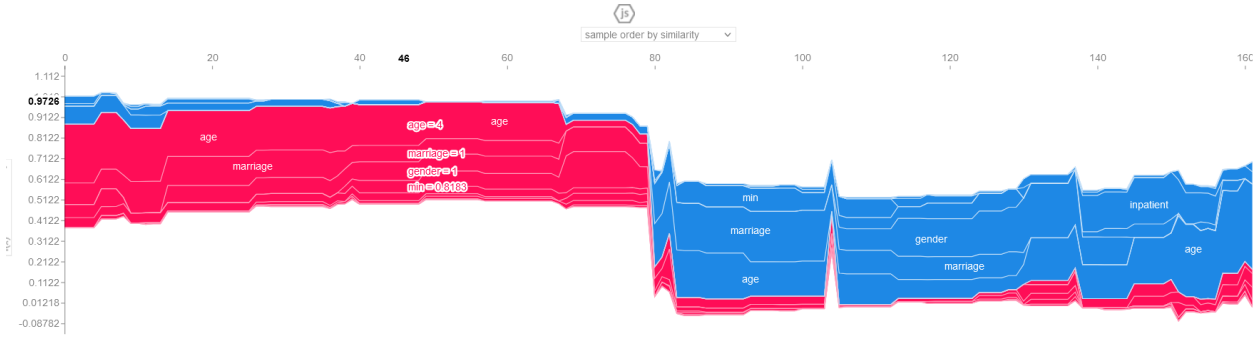

Figure 24. The effect of each feature in the whole dataset of the second model

By observing the diagram of Fig. 25 one can be confused from the distribution of the value of age and its effect in the model prediction. On the other hand, it is clear that the model believes that a person admitted to hospital, a woman, a single, higher educated person with lower min and max value have a tendency towards bipolar in contrast with unipolar depression and vice versa.

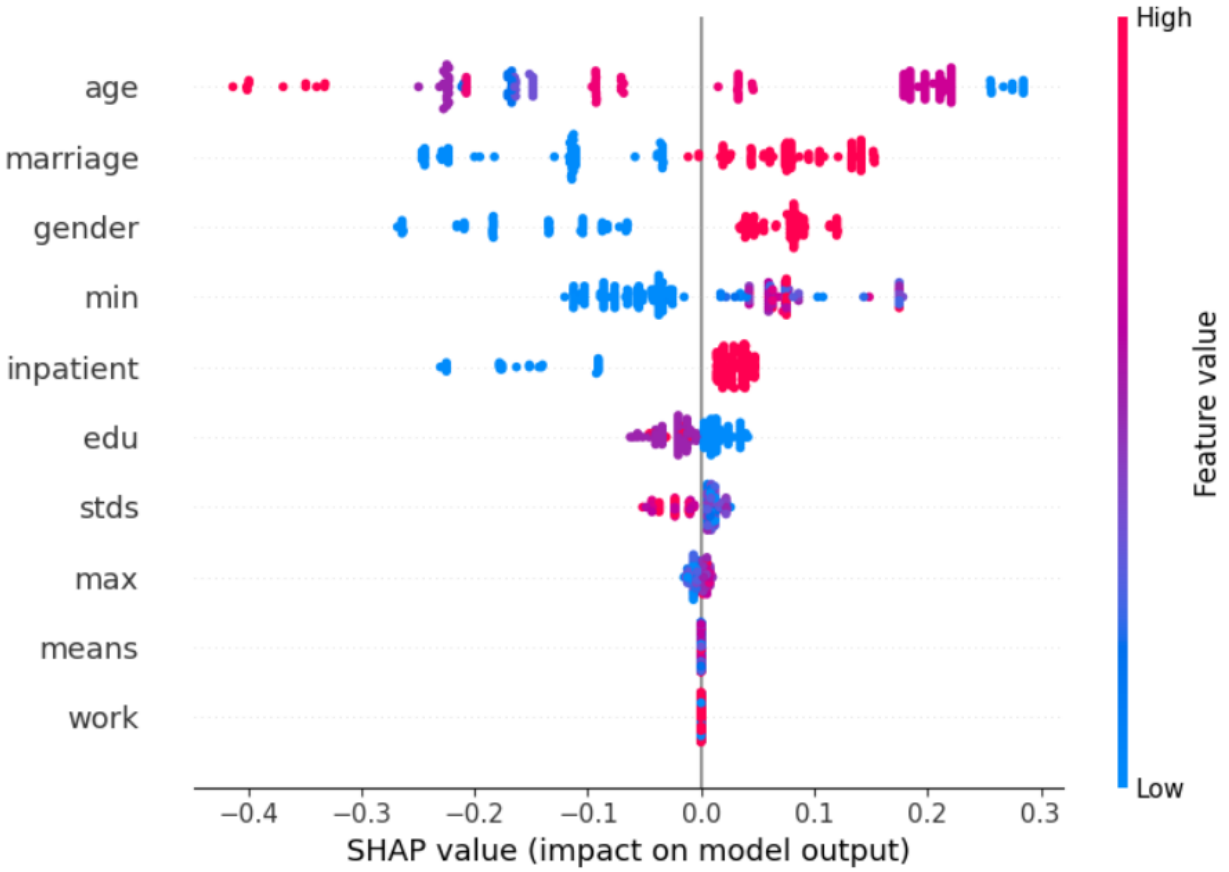

Figure 25. The volume of the effect of each value of the feature in second model

7 Conclusion

After re-doing most of the parts of this assignment we spotted some differences that highlight the importance of proper dataset handling so as to not result in a false outcome. Neither of our neural networks manage to categorize the dataset in two classes for our two classification problems leading us to rely on the machine learning approaches. It's a pity not to utilize and implement the research for multimodal explainability, but it would be inappropriate to explain a model with close to random predictions. Needless to say, if a problem can be addressed in lower capacity models then there is no need of adding more complexity and computational cost. We presented some cases of our explainable models and demonstrate how reasoning can be gain for the prediction of the models. We believe it is vital for medical professionals to be provided with explainable models so as to determine whether or not they can be trusted and in which grade.

8 References/Bibliography

[1] Littlewort G et al (2011) The computer expression recognition toolbox (cert) 298–305 (IEEE)

[2] Baltrusaitis T, Zadeh A, Lim YC, Morency L-P (2018) Openface 2.0: Facial behavior analysis toolkit 59–66 (IEEE)

[3] Wang Q, Yang H, Yu Y (2018) Facial expression video analysis for depression detection in Chinese patients. J Vis Commun Image Represent 57:228–233. Article: https://www.sciencedirect.com/science/article/abs/pii/S1047320318302761?via%3Dihub

[4] Cootes T, Edwards G, Taylor C (2001) Robust real-time periodic motion detection, analysis, and applications. IEEE Trans Patt Analy Mach Intell 23(6):681–685. Article: https://ieeexplore.ieee.org/document/927467

[5] Cummins N, Epps J, Breakspear M, Goecke R (2011) An investigation of depressed speech detection: features and normalization

[6] Cummins N, Sethu V, Epps J, Schnieder S, Krajewski J (2015) Analysis of acoustic space variability in speech affected by depression. Article: https://www.sciencedirect.com/science/article/abs/pii/S0167639315000989?via%3Dihub

[7] Thin Nguyen, Dinh Phung, Bob Dao, Svetha Venkatesh, and M. BerkMichael Berk, "Affective and content analysis of online depression communities," IEEE Transactions on Affective Computing, vol. 5, no. 3, pp. 217–226, 2014.

[8] Alghowinem S. et al (2016) Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors. IEEE Trans Affect Comput 9(4):478–490 Article: https://ieeexplore.ieee.org/document/7763752

[9] Joshi, G., Walambe, R., & Kotecha, K. (2021). A review on explainability in multimodal deep neural nets. *IEEE Access*, *9*, 59800-59821. Article: https://link.springer.com/article/10.1007/s11042-022-12315-2#ref-link-section-d2929429e949

[10] Guarrasi, V., Tronchin, L., Albano, D., Faiella, E., Fazzini, D., Santucci, D., & Soda, P. (2022). Multimodal explainability via latent shift applied to COVID-19 stratification. arXiv preprint arXiv:2212.14084.