USC SURE Final Report

Geffen Cooper

Advisor: Mohammad Soleymani

June - September 2021

1 Project Overview

This section explains the purpose and goals of this project

1.1 Project Topic: Assessing Psychomotor Retardation

Much of the literature in computational mental health seeks to directly assess the severity of psychiatric disorders such as Major Depressive Disorder (MDD). However, due to heterogeneous nature of these disorders, it seems that the analysis of individual symptoms might be a more effective approach [1]. Accordingly, this project instead seeks to assess psychomotor retardation (PMR), a *specific symptom* present in multiple psychiatric disorders (most commonly in MDD).

Psychomotor Retardation (PMR): a slowing down or inhibition of mental and physical activity, manifest as slow speech with long pauses before answers, slowness in thinking, and slow body movements.

1.2 Research Question

How can psychomotor retardation effectively be assessed and tracked using audio and visual data?

1.2.1 Importance

- A noninvasive method for assessing the presence or severity of PMR may provide insight into higher level disorders such as MDD.
- The ability to track PMR severity over time may help clinicians assess the efficacy of different treatments for related psychiatric disorders.

1.3 Goals

1. Identify potential behavioral biomarkers of PMR
2. Build a model that predicts the presence and/or the severity of PMR

* *These goals are in the context of the audiovisual data from AVEC 2019 DDS*

2 Identifying Potential Behavioral Biomarkers of PMR

This section describes the first phase of the project

2.1 Description

The following are a set of audio and visual features that have been utilized in literature associated with detection and diagnosis (severity) of depression. These features will be examined as potential candidates for detecting or measuring the condition of psychomotor retardation by comparing these values in healthy and depressed individuals over time to see if there is a correlation.

2.2 Standard Features that are Popular in the Literature

2.2.1 Audio

- Formant Frequencies (Filter Features) [2], [3], [4], [5], [6]
  - formants are associated with the frequency response of the vocal tract i.e. based on the shape of the vocal tract, certain frequencies will be intensified or resonate. Tracking formant frequencies can serve as a way to track the shape/dynamics of the vocal tract. Measures that potentially correlate with depression are:
    - **F1 and F2 mean, variance, standard deviation, and range**
- MFCCs, and delta MFCCs (Spectral Features) [2], [3], [4], [7], [5]
  - MFCCs and delta MFCCs are commonly used audio features that can provide relevant spectral structure information. Measures that potentially correlate with depression are:
    - **Values of certain coefficients**
- Prosodic Features [4], [8], [9], [5]
  - These features are associated with longer term variations in speech such as rhythm, stress, and intonation. Measures that potentially correlate with depression are:
    - **f0 variability and range. Intensity variability. Energy velocity, speech rate, pause rate, and total pause time**
- Source Features [4], [10], [9], [5]
  - These features are associated with voice production and voice quality. Measures that potentially correlate with depression are:
    - **Jitter, shimmer, and HNR**
- Most of these features are collectivized under the Extended Geneva Minimalistic Acoustic Parameter Set [11]

2.2.2 Video

- Facial expressions, movement, and pose based on facial action units [12], [13], [14]
  - these features can provide insight into emotional state and reactivity. Measures that potentially correlate with depression are:
    - **intensities of facial action units (e.g. AU 12,14,15,24) over time, amplitude and velocity of head motion, vertical head gaze, vertical eye gaze, smile intensity, smile duration**

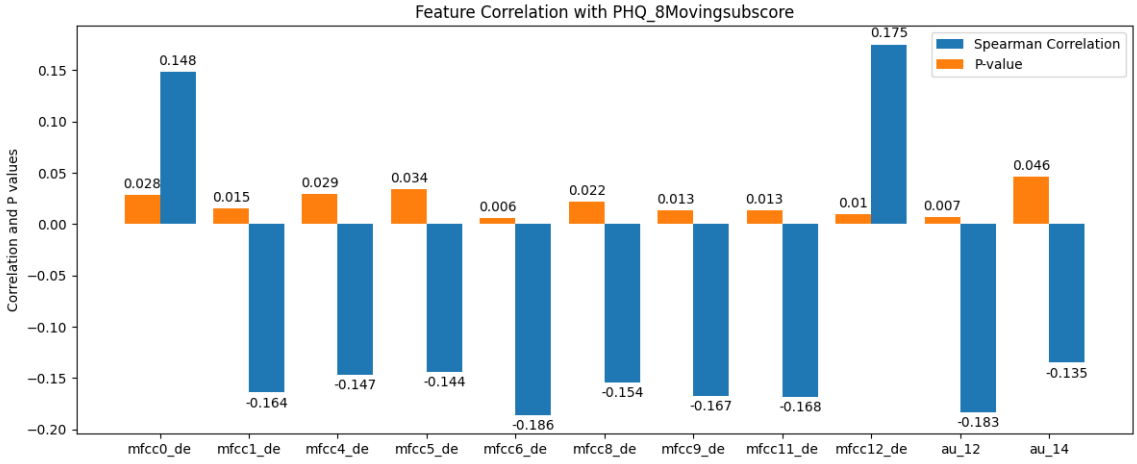
3 Correlation Results

This section presents the results of the correlation analysis

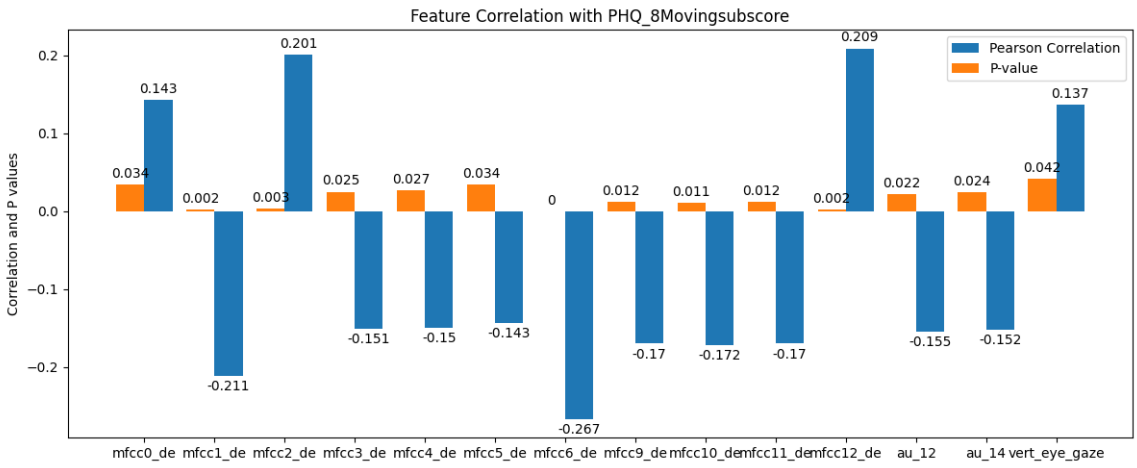
3.1 Description

The following show the correlation results using the audio and visual features from the AVEC 2019 DDS data. There are two sets of results. The first uses features from the baseline patient responses. The second uses features from the segments of patient responses where there are vowel transitions (e.g. /ai/, /oi/) with the goal of using potentially more salient regions with regard to psychomotor retardation. Only features that have $p < 0.05$ are shown for 'PHQ_Total' and 'PHQ8_Moving' subscores. The complete results are on GitHub [1]. Both Pearson and Spearman correlations are shown.

3.2 Baseline Responses



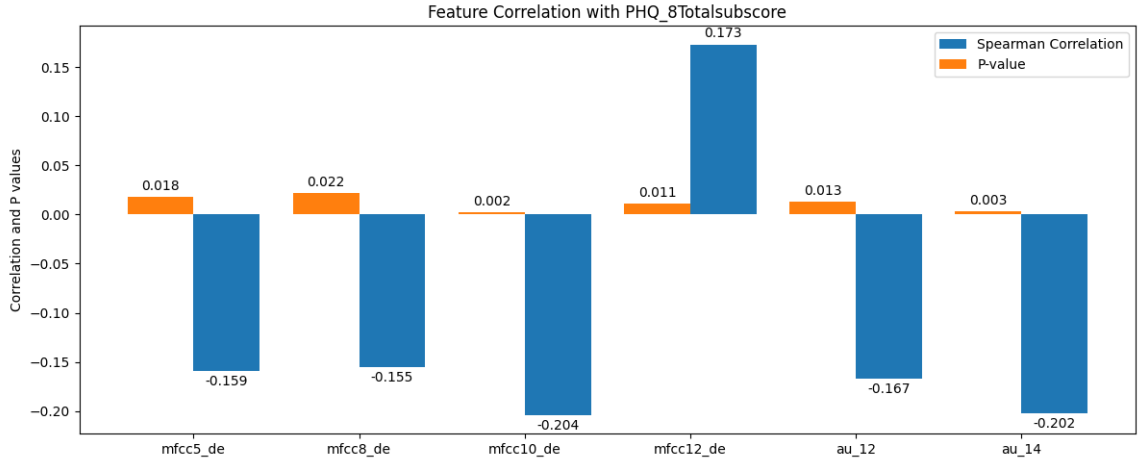
(a) Spearman



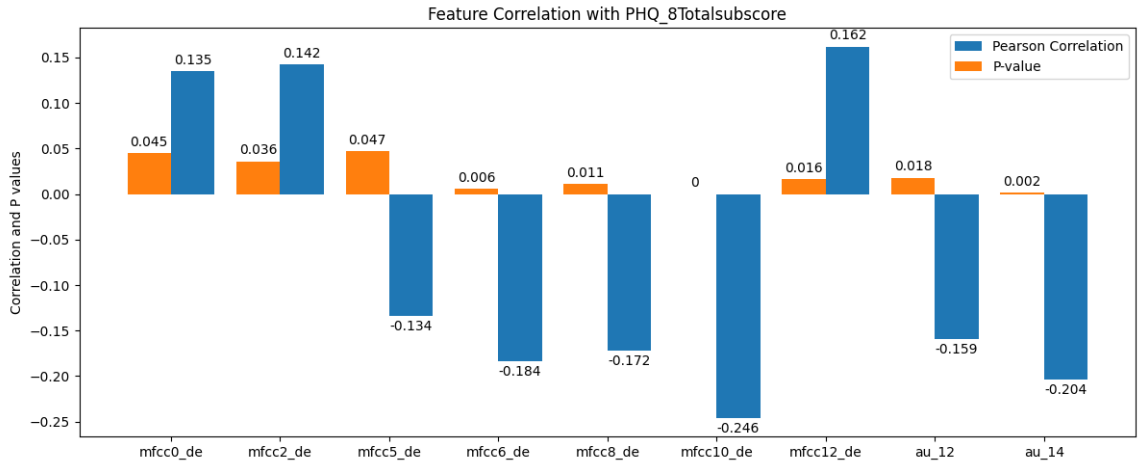
(b) Pearson

Figure 1: Correlation and P Values for PHQ8_Moving Subscore

[1] https://github.com/geffencooper/PMR_correlations/tree/dev/data



(a) Spearman



(b) Pearson

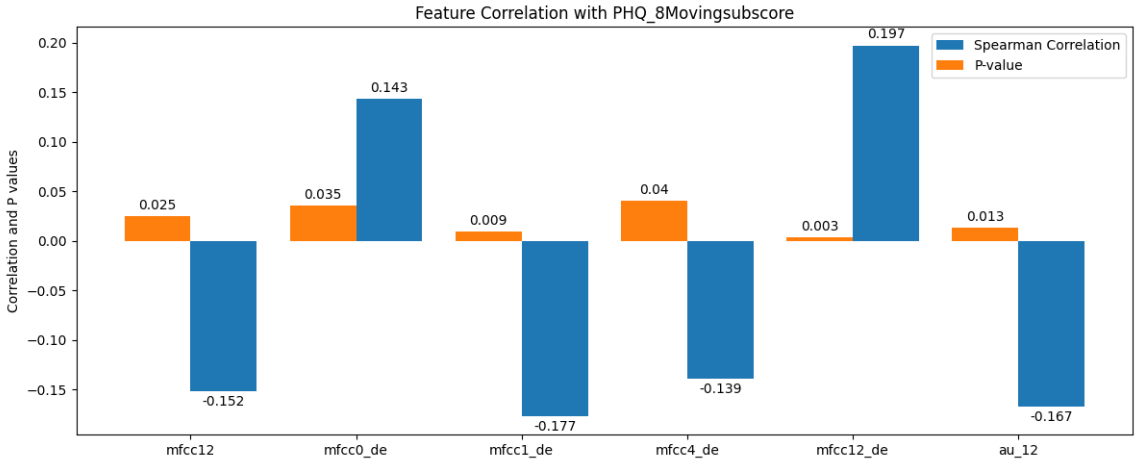
Figure 2: Correlation and P Values for PHQ8_Total Subscore

3.2.1 Baseline Responses Observations

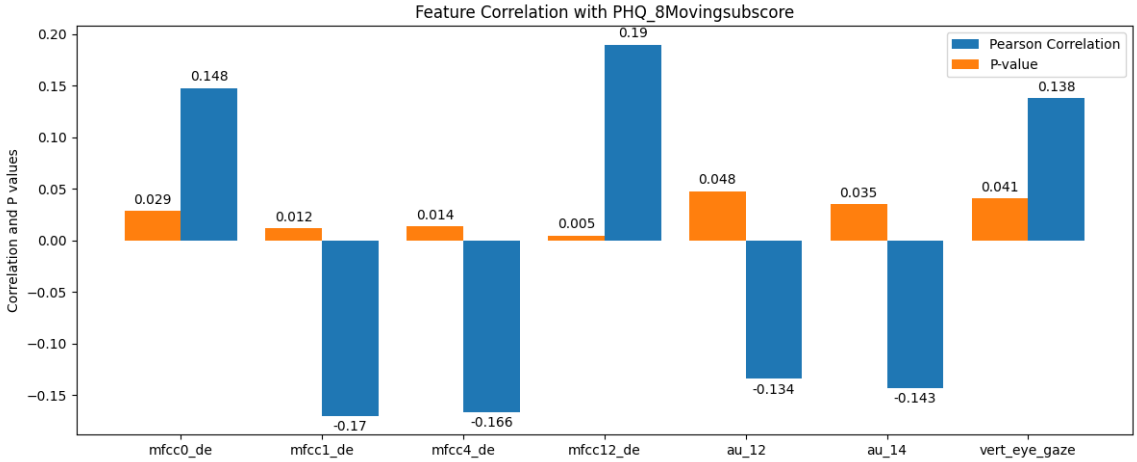
The delta MFCCs tended to correlate with each depression subscore but these can be difficult to interpret and it does not seem that a specific one is much more notable than the others. AU12 and AU14 negatively correlated with the moving and total subscores. This makes sense since we expect individuals with MDD to smile less and have less facial movement in general. In addition, vertical eye gaze angle has a positive correlation with the PHQ8_Moving subscore which may be interpreted as the patient looking down.

3.3 Segmented Responses Based on Vowel Transitions

The vowel transitions used were [“igh”,“oi”,“ai”,“ie”,“ia”,“ime”,“ike”,“ua”]. This selection is somewhat arbitrary but the assumption is that individuals with psychomotor retardation may have less motor articulation during speech production. Ideally, this effect of reduced articulation should be observed through reduced formant transitions [6]. However, there were no new significant correlations as shown by the results. This may be due to the usage of conversational speech rather than selected phrases.

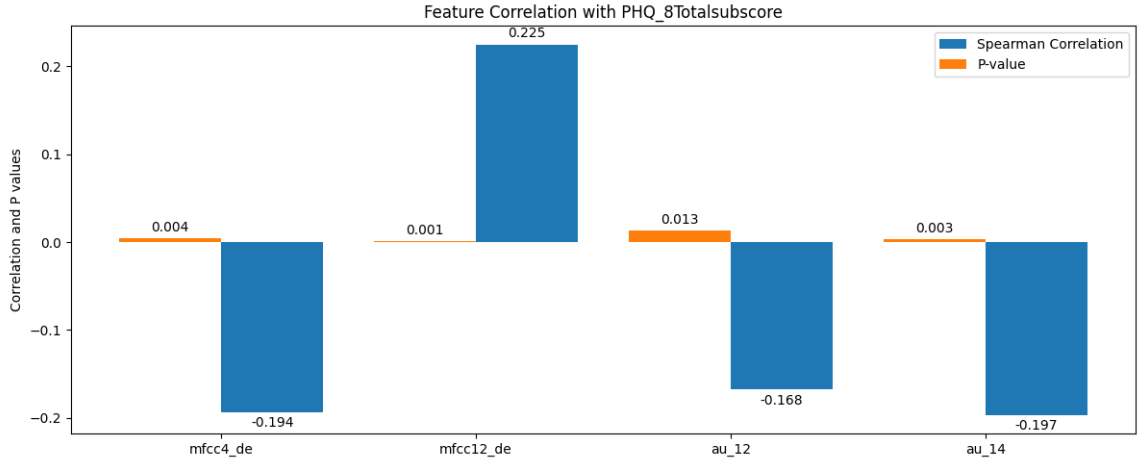


(a) Spearman

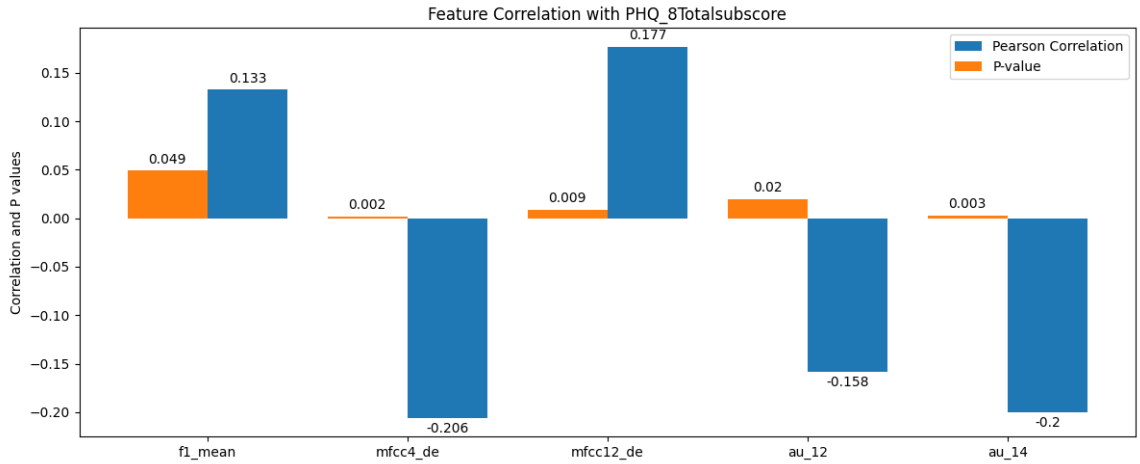


(b) Pearson

Figure 3: Correlation and P Values for PHQ8_Moving Subscore



(a) Spearman



(b) Pearson

Figure 4: Correlation and P Values for PHQ8_Total Subscore

3.4 Overall Observations

No directly interpretable acoustic feature (i.e. features other than MFCCs) had a significant correlation with the shown depression subscores. Other PHQ8 depression subscores did have correlations with the other acoustic features but they are not associated with Psychomotor Retardation. For example, the PHQ8_Failure subscore had a negative correlation with the second formant variance/standard deviation. Overall, the visual and MFCC features seemed to be more salient in this analysis. For any conclusions to be drawn other datasets across different disorders with psychomotor retardation as a symptom should be examined.

4 Predicting the Presence/Severity of PMR

This section explains phase two of this project

4.1 Taking a Deep Learning Approach

At least within the context of MDD, there currently does not seem to be a definitive choice of biomarkers for identifying the disorder prescence/severity[1]. This coupled with the recent success and rapid growth of deep learning has prompted the use of deep learning methods for depression recognition [15]. Thus, in this project a deep learning model was used to try and predict the presence/severity of PMR.

4.1.1 Transfer Learning

One of the main challenges of using deep learning in a medical setting is the limited availability of training data. A common technique to overcome this challenge is transfer learning. In general, transfer learning refers to the idea that knowledge gained from solving one task can be applied to a different but related task.

4.2 Speech Rate Detection

In this project, the model was first trained on the proxy task of speech rate detection. This task was motivated by the fact that a common indicator of PMR is slowed speech and body movements.

4.2.1 Dataset

The dataset used for this task was a subset of the Common Voice dataset [2] from kaggle [3] which contains a diverse set of short speech segments that are 2-5 seconds in duration. There were 200K training samples, 4K validation samples, and 4K test samples.

4.2.2 Preprocessing

These speech samples were evenly split into three classes (normal, sped up, slowed down) and preprocessed by time stretching using a random factor. Specifically, the samples for the *sped up* class were sped up by a random factor between 1.1 and 1.65, whereas the samples for the *slowed down* class were slowed down by a factor between 0.35 and 0.9. The samples for the *normal* class were left alone. To make these alterations more comparable to actual speech, variations in the time stretching factor could have been applied throughout each segment rather than using a single factor for the whole segment. However, due to time constraints this was not attempted.

4.2.3 Augmentation

In order to make this task more difficult and to prevent overfitting, random data augmentations were applied to the audio signals after preprocessing. These included the addition of Gaussian noise and pitch shifting by random amounts [4].

4.2.4 Input Features and Architecture

Rather than using raw audio as the input for speech pace detection, MFCCs and delta MFCCs were extracted using openSMILE [16]. These not only provide useful spectral structure information, but also condense the sequence length. RNNs are commonly used for processing time series or sequential data and so a GRU was used for this model. The pipeline and architecture can be seen in Figure 5. During training each input matrix of features was also normalized through time by subtracting the mean and dividing by the standard deviation.

[2] https://commonvoice.mozilla.org/en/datasets
[3] https://www.kaggle.com/mozillaorg/common-voice
[4] https://github.com/iver56/audiomentations

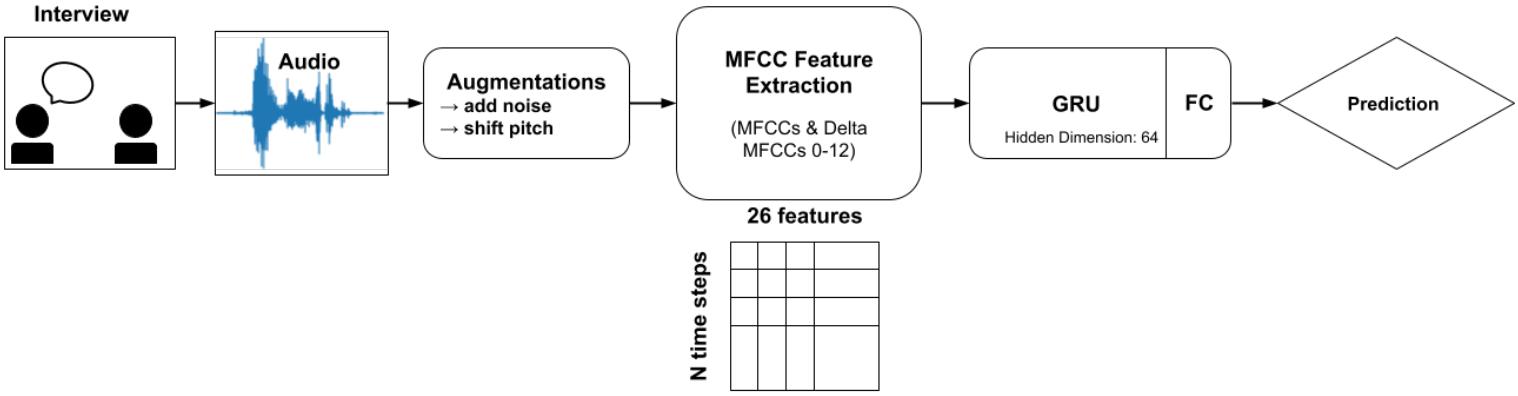


Figure 5: Speech Pace Detection Model

4.2.5 Results

This model was able to get a test accuracy of 88% when trained with RMSprop with zero initialization for the hidden state.

4.3 Predicting PMR

To predict the prescense/severity of PMR the pretrained speech detection model was expanded to use visual features (Facial Action Units and Head Pose). This model can be seen in Figure 6 and was used to predict the presence of PMR (binary classification) as well as the severity (regression). The audio and visual modalities were processed separately using GRUs. The last set of output features from each GRU (many to one) was concatenated (fusion) and then finally passed through a fully connected layer for prediction.

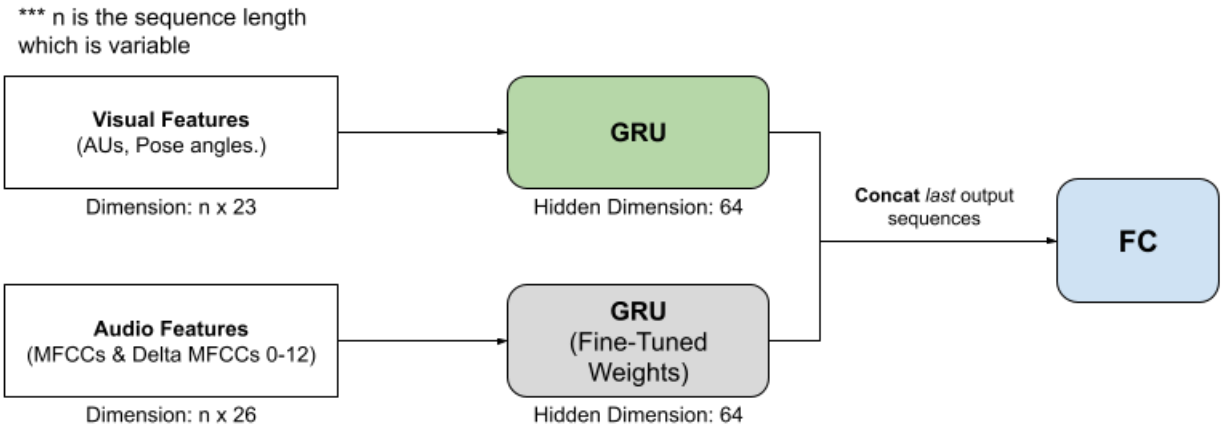


Figure 6: PMR Prediction Model

4.3.1 Dataset

The dataset used for this task was the AVEC 2019 DDS data [17] (pre-extracted features). The labels used to assess psychomotor retardation were the PHQ8_Moving subscores.

4.3.2 Class Imbalance

One of the major challenges of working with this dataset is the severe class imbalance in the PHQ8_Moving sub cores

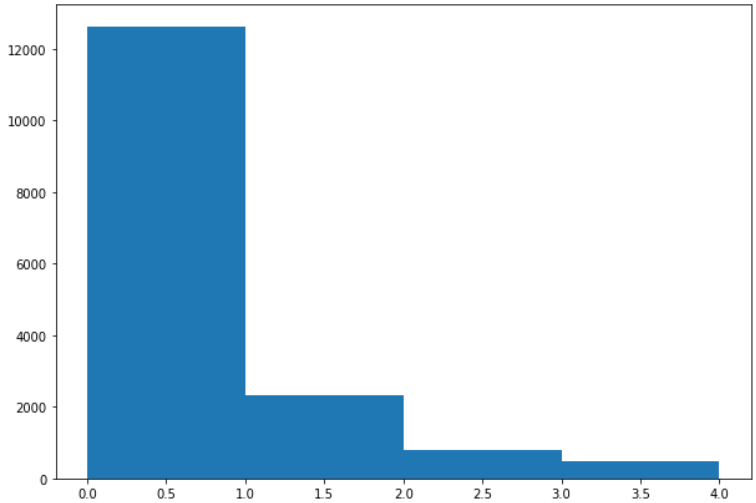


Figure 7: Class Distribution for PHQ8_Moving Subcore (0,1,2,3)

As seen in Figure 7, the vast majority of samples are class 0 which represents a score (severity) of 0. Without accounting for this unbalanced distribution, the model tends to learn to just predict class 0 almost every time since this will result in a relatively low loss. Three methods were applied to attempt to mitigate this issue.

1. In method one, a sampler was used to rebalance the class distribution per batch [5] as shown in Figure 8.
2. In method two, class weights were added to the loss function so that there was a larger penalty for incorrectly classifying the minority class.
3. In method three, artificial audio and visual features were generated using SMOTE for the minority classes to rebalance the class distribution.

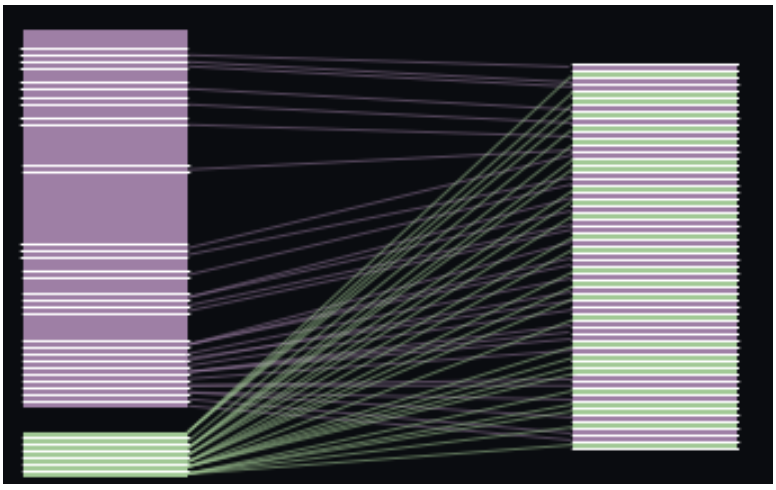


Figure 8: Rebalancing the Class distribution for two classes

[5] https://github.com/ufoym/imbalanced-dataset-sampler

Despite each of these methods, none of the model variants were able to predict the presence or severity of PMR with accuracy better than random on the validation set. When looking at the training and validation loss curves, the model tends to overfit. As a result, dropout and L2 regularization were also employed but had negligible effects.

5 Conclusion

This project seeked to build a model for predicting the presence and severity of psychomotor retardation. Unfortunately, no significant results were achieved with regard to classification accuracy or mean squared error for predicting the PHQ_Moving subscore. This may be due to an oversimplified model, insufficient training data, or invalid assumptions with regard to the input features. For a more complete assessment of this task:

- other architectures should be explored including a transformer based model
- a more realistic approach for speech pace detection could have been attempted as mentioned in section 4.2.2

6 Acknowledgements

Despite the insignificant results, I enjoyed working on this project and it was a valuable experience to learn to work with time series data, multiple modalities, and imbalanced classes. I want to thank Professor Soleymani for mentoring me through this project as well as Trang and Leili for answering my questions and providing advice.

7 Code

All the code for this project is located at:

- Correlation Analysis: https://github.com/geffencooper/PMR_correlations/tree/dev/data
- Deep Learning Model: https://github.com/geffencooper/PMR_prediction

References

[1] E. I. Fried and R. M. Nesse, "Depression sum-scores don't add up: why analyzing specific depression symptoms is essential," *BMC Medicine*, vol. 13, p. 72, Apr 2015.

[2] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, AVEC '13, (New York, NY, USA), p. 41–48, Association for Computing Machinery, 2013.

[3] T. Quatieri, J. Williamson, A. Lammert, K. Heaton, and J. Palmer, "Noninvasive biomarkers of neurobehavioral performance," 2020.

[4] D. Low, K. Bentley, and S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope Investigative Otolaryngology*, vol. 5, 01 2020.

[5] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[6] A. J. Flint, S. E. Black, I. Campbell-Taylor, G. F. Gailey, and C. Levinton, "Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression," *Journal of Psychiatric Research*, vol. 27, no. 3, pp. 309–319, 1993.

[7] T. Taguchi, H. Tachikawa, K. Nemoto, M. Suzuki, T. Nagano, R. Tachibana, M. Nishimura, and T. Arai, "Major depressive disorder discrimination using vocal acoustic features," *Journal of Affective Disorders*, vol. 225, 08 2017.

[8] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, "Vocal acoustic biomarkers of depression severity and treatment response," *Biological Psychiatry*, vol. 72, no. 7, pp. 580–587, 2012. Novel Pharmacotherapies for Depression.

[9] T. Quatieri and N. Malyska, "Vocal-source biomarkers for depression: A link to psychomotor activity," in *INTERSPEECH*, 2012.

[10] A. Ozdas, R. Shiavi, S. Silverman, M. Silverman, and D. Wilkes, "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 9, pp. 1530–1540, 2004.

[11] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[12] S. Scherer, G. Stratou, G. Lucas, M. Mahmoud, J. Boberg, J. Gratch, A. S. Rizzo, and L.-P. Morency, "Automatic audiovisual behavior descriptors for psychological disorder analysis," *Image and Vision Computing*, vol. 32, no. 10, pp. 648–658, 2014. Best of Automatic Face and Gesture Recognition 2013.

[13] J. Joshi, R. Goecke, G. Parker, and M. Breakspear, "Can body expressions contribute to automatic depression analysis?," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–7, 2013.

[14] J. Girard, J. Cohn, M. Mahoor, S. Mavadati, Z. Hammal, and D. P. Rosenwald, "Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses," *Image and vision computing*, vol. 32 10, pp. 641–647, 2014.

[15] L. He, M. Niu, P. Tiwari, P. Marttinen, R. Su, J. Jiang, C. Guo, H. Wang, S. Ding, Z. Wang, W. Dang, and X. Pan, "Deep learning for depression recognition with audiovisual cues: A review," 2021.

[16] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, (New York, NY, USA), p. 1459–1462, Association for Computing Machinery, 2010.

[17] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani, and M. Pantic, "Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition," 2019.