MINI PROJECT 1
COGS 209 Scientific Data Analysis and Statistical Learning

Depression Identification using Continuous Motor Activity

Lead Author: Sowmya Manojna Narasimha

Co-Authors: Morgan Fitzgerald, Michaela Cullum-Doyle, Vanessa Martin

University of California, San Diego

Contents

1 Introduction

1.1 Background

The global incidence of depression is on the rise (Marrie et al., 2019). This alarming trend highlights the urgent need for new and effective tools to detect and treat depression. Early detection and treatment of depression are crucial for improving outcomes and preventing further complications. Fortunately, recent advancements in technology have opened up new avenues for detecting and tracking depression using digital tools. One promising idea is to use motor activity data to classify the depressive state of individuals (Aminifar et al., 2021), (Garcia-Ceja et al., 2018). This approach is exciting as it has the potential to overcome the limitations of canonical depressive-symptom surveys and provide a more objective measure of an individual's depressive state.

In this study, we aim to investigate the potential of motor activity as a predictor of depressive states using motor activity (actigraph data) from (Garcia-Ceja et al., 2018). We will leverage several classifications models including Logistic Regression, Random Forests, K-Nearest Neighbors, Support Vector Machines, and Multi-Layered Perceptrons, and compare their performance in predicting depression based on motor activity. We will include sex and age as covariates in our analysis, as they may impact motor activity and have implications for the accuracy of our models. We hypothesize that the random forest algorithm will outperform the linear regression model in classifying depressive state based on motor activity data (H1). We also expect that age and sex will have a significant impact on motor activity and that including them as covariates in the logistic regression model will improve its performance in depression classification (H2). Finally, we predict that the model will perform better on males compared to females due to potential differences in motor activity patterns between sexes (H3) and on individuals with severe depression compared to mild/moderate depression as their motor activity may be more significantly impacted by their depressive state (H4). Overall, this study has important implications for the development of accurate and reliable tools for detecting and tracking depression using machine learning algorithms and digital tools.

2 Methods

2.1 Data

2.1.1 Description

For all of our analysis, we used the *Depresjon* dataset presented in (Garcia-Ceja et al., 2018). The dataset consists of movement activity, actigraph data data recorded using a wrist watch (actiwatch). The actiwatch measures activity as a function of the duration, amount and intensity of movement in all directions, using a piezo-electric accelerometer. The total activity count was measured in 1 minute periods.

The actigraph data was measured from 23 patients who were diagnosed with depression and 32 control subjects. In addition to the actigraph data from each of the participants, meta data about each participant was also made available. The fields present in the meta data are as follows:

1. `number`: Patient Identifier (Control or Condition)
2. `days`: Number of days of measurements
3. `gender`: Gender indicator (1: Females, 2: Males)
4. `age`: Age in age groups
5. `afftype`: Type of depression (1: Bipolar Depression II, 2: Unipolar Depression, 3" Bipolar Depression I)
6. `melanch`: Melancholia Identifier (1: Melancholia, 2: No melancholia)
7. `inpatient`: Inpatient Identifier (1: Inpatient, 2: Outpatient)
8. `edu`: Education-level grouped in years

9. `marriage`: Marriage/Relationship Indicator (1: Married/Cohabiting, 2: Single)
10. `work`: Work Indicator: (1: Working/Studying, 2: Unemployed/On a Sick Leave/Receiving Pension)
11. `madrs1`: MADRS score when measurement started
12. `madrs2`: MADRS score when measurement stopped

2.1.2 Preporcessing

2.1.2.1 Meta Data Imputation The meta data provided with the actigraph data, had missing entries. The missing value imputations were performed using a K-Nearest Neighbors based imputation. The neighbors were weighted based on their distance for the imputation and points that were closer were given higher weightage.

2.1.2.2 Actigraph Data Preprocessing Actigraph data from each participant was obtained. The number of samples varied for each participant as the days across which the data was collected varied. For each participant, the data was split across days and only the data where motor activity was available for entire duration (1440 minutes in a day) was used for further analysis. In order to overcome the limitations caused due to the small sample size, each day was considered as a datapoint for all subsequent analysis. In addition to considering only days where motor activity was available for the whole duration, data from only the days with average activity greater than a threshold activity were considered. This is explained using graphics in Figure 1.

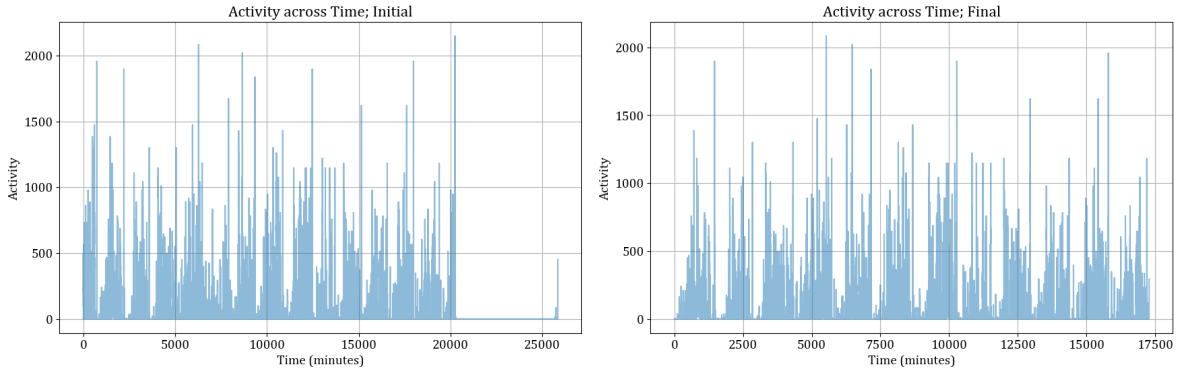

Figure 1: Motor Activity of Condition Participant 20. Note the extended duration of no activity towards the end in the graph on left. Final processed actigraph data on the right.

A consolidated dataset was obtained by stacking the data of each day for every participant across both control and condition groups. Another instance of the data was created, where the activity data was normalized for each participant, to avoid any scaling issues arising due to individual differences.

2.1.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was performed on the meta data. Pair plots and correlations were analyzed for any trends. Some observations that we can draw from the correlation plot in Figure 2 is as follows:

- `work` and (`melanch`, all `madrs` scores) are highly positively correlated. This is particularly interesting because higher values of `work` (2), indicates that the participant is unemployed/on a sick leave/receives pension. A positive correlation implies that people who are unemployed/on a sick leave/receive pension, tend to be more depressed than people who are currently working or studying.
- `marriage` and (`melanch`, all `madrs` scores) are highly positively correlated. A higher value of `marriage` (2) indicates that the participant is single. Hence, the positive correlation implies that single people tend to be more depressed.

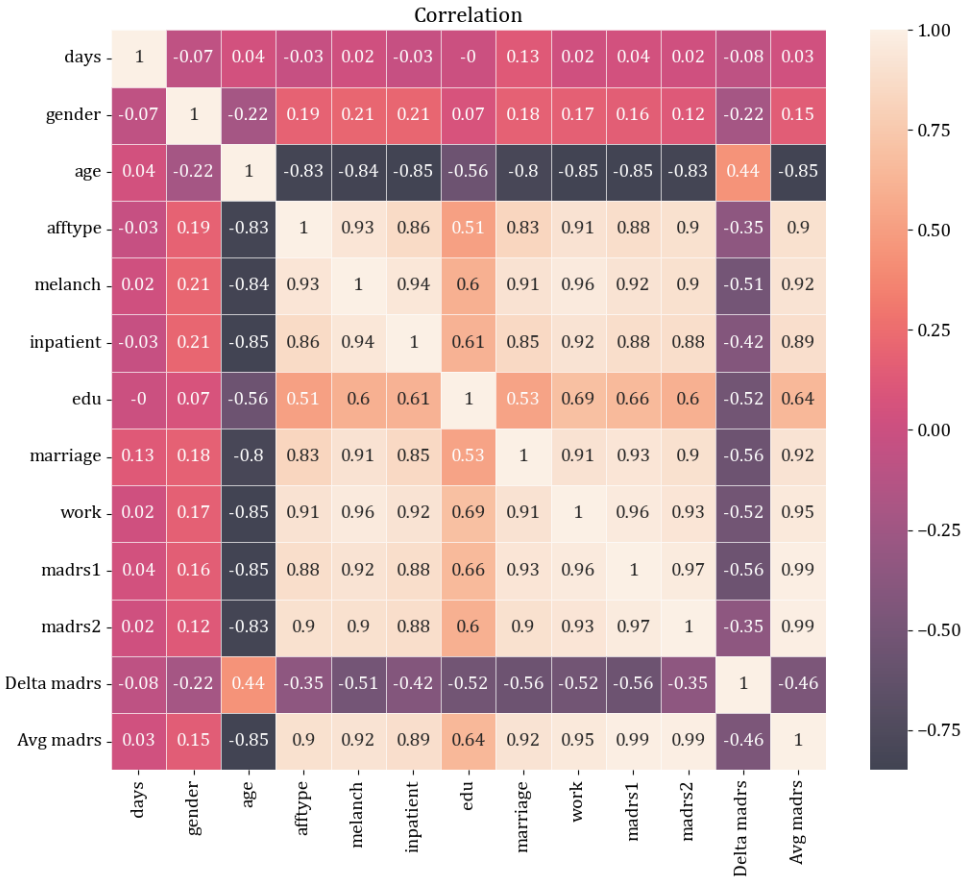

Figure 2: Correlation Analysis using entries in the meta data.

The pair plots were plotted to get a better understanding of the distribution of the values in the meta data. From the pair shown in Figure 3, we can see that the data is lightly imbalanced in most classes. Hence, for all further analysis, the $F1$ Scoring metric will be used for model evaluation.

2.2 Classification Models

In order to address all of our hypothesis, we used classification models from `sklearn`. Parameters that best explained the data was identified using Grid Search, 5-fold cross validation and $F1$ Score as the metric. The classification models considered are as follows:

- Logistic Regression (LR)
- Logistic Regression (LR), with Standard Scaler
- Random Forest (RF)
- Random Forest (RF), with Standard Scaler
- K-Nearest Neighbors Classifier (KNN)
- K-Nearest Neighbors Classifier (KNN), with Standard Scaler
- Support Vector Machine Classifier (SVC)
- Support Vector Machine Classifier (SVC), with Standard Scaler
- Multi Layered Perceptron Classifier (MLP)
- Multi Layered Perceptron Classifier (MLP), with Standard Scaler

These models were trained on both un-normalized and normalized data. The parameter for each case that resulted in the best mean $F1$ score were saved and used on the validation dataset.

The hyperparameters considered for each model (normal and standard scaled) is as follows:

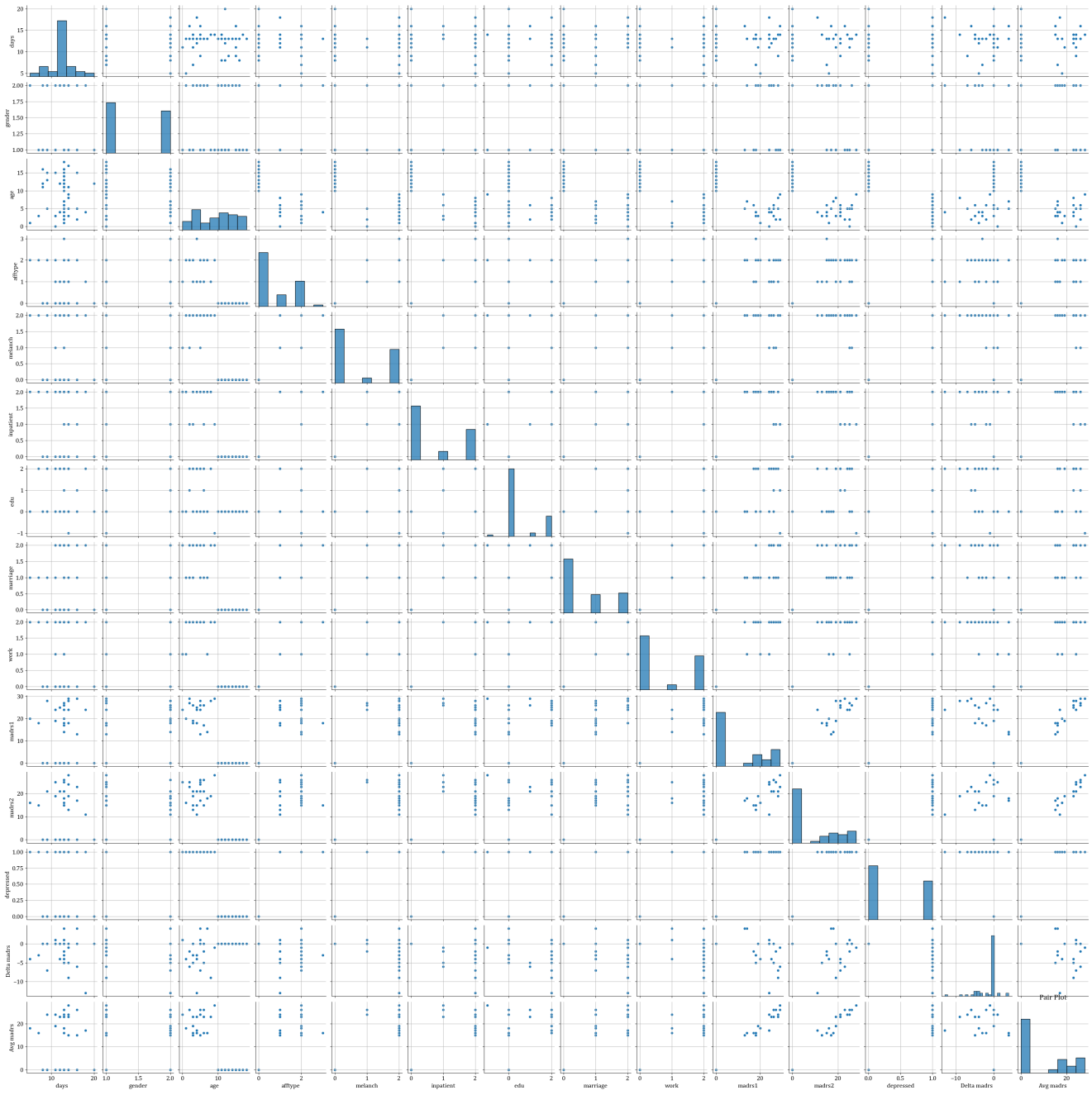

Figure 3: Pair Plot of values in the meta data. Plot is made smaller for ease of representation, but is of high resolution and can be zoom in for more details.

1. LR: `penalty`=[`l1`, `l2`, `elasticnet`, `None`], `C`=[0.1, 0.5, 1, 10, 50], `solver`=[`liblinear`,`saga`]
2. RF: `n_estimators`=[50, 100, 150], `min_samples_split`=[2, 4], `max_features`=[`sqrt`, `llog2`]
3. KNN: `n_neighbors`=[3, 5, 10, 15], `weights`=[`uniform`, `distance`]
4. SVC: `C`=[0.1, 0.5, 1, 10, 50], `kernel`=[`linear`, `poly`, `rbf`, `sigmoid`], `degree`=[3, 5, 10, 20], `decision_function_shape`=[`ovo`, `ovr`], `gamma`=[`auto`, `scale`]
5. MLP: `hidden_layer_sizes` = [(100,), (50,), (50,10)], `activation` = [`relu`, `linear`], `alpha` = [1e-6, 1e-4, 1e-2], `learning_rate` = [`constant`, `invscaling`, `adaptive`]

3 Results & Comparisons

The performance of the models on the training data, with the best parameter set returned is as shown below:

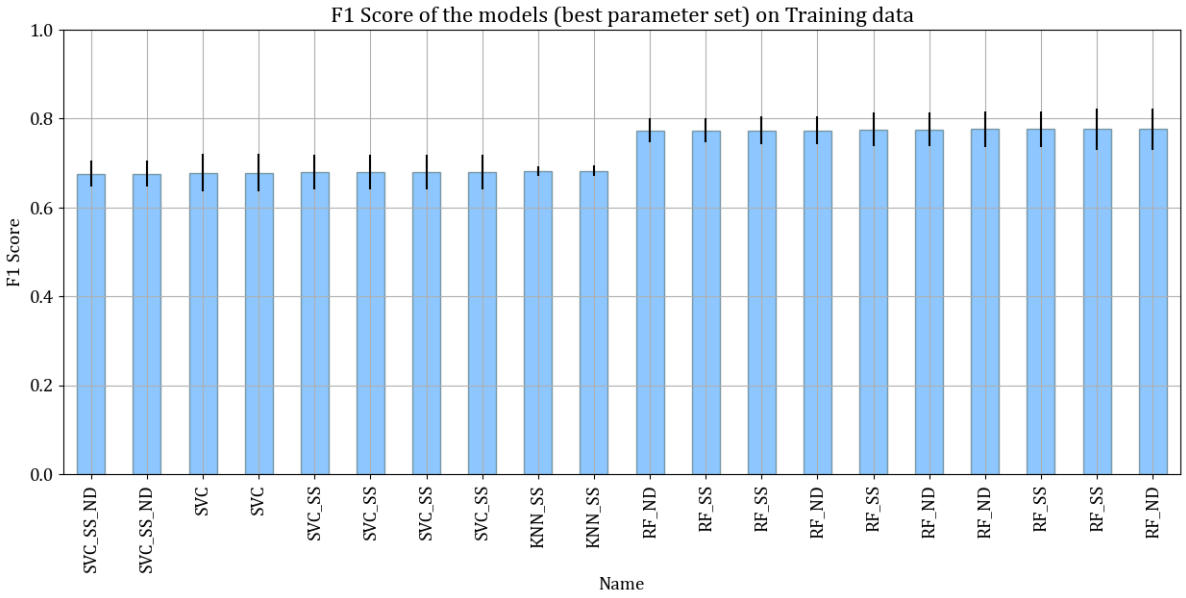

Figure 4: F1 Scores of the top 20 Model, Parameter combinations returned by Grid Search CV.

From Figure 4, the best performing models on the dataset are Random Forest Models, followed by KNN Classifiers and SVM Classifiers. For ease of representation, in all the subsequent results "`_ND`" indicates that the data is trained on the patient-wise normalized data; "`_SS`" indicates that the whole data was standard scaled before applying the model.

3.1 Logistic Regression

The top 5 model, tuned with parameters obtained using Grid Search are as follows:

Name	C	Regularization	Solver	F1	Accuracy
LR_ND	0.5	l1	liblinear	0.77	0.83
LR_SS	0.1	l2	liblinear	0.75	0.78
LR	50	l2	saga	0.75	0.79
LR	10	l2	saga	0.75	0.79
LR	10	None	saga	0.75	0.79

Table 1: Top 5 performances of the Logistic Regression Classifier model, sorted based on validation F1 scores, for different parameter values.

Interpretations based on the best model parameters:

1. The best model has a relatively low `C` value. As `C` is the inverse of regularization coefficient in `sklearn`, this implies that the best model utilizes regularization to improve generalization.
2. The regularization used by the best model is `l1` regularization. The second best model uses the `l2` regularization, but in order to achieve the same level of performance as the best model, it uses stronger regularization.
3. The solver that gives the best F1 score is `liblinear`. It has been shown in (Fan et al., 2008) that `liblinear` is a very good choice for smaller datasets, in practice.

3.2 Random Forest

The top 5 model, tuned with parameters obtained using Grid Search are as follows:

Name	Max Features	Min Samples Split	# Estimators	F1	Accuracy
RF_ND	sqrt	2	150	0.89	0.90
RF_ND	sqrt	2	50	0.89	0.90
RF_ND	sqrt	2	100	0.88	0.90
RF_ND	sqrt	4	100	0.88	0.90
RF_ND	sqrt	4	50	0.87	0.89

Table 2: Top 5 performances of the Random Forest Classifier model, sorted based on validation F1 scores, for different parameter values.

Interpretations based on the best model parameters:

1. `min_samples_split` of 2 gives better accuracy than of 4. As this parameter determines the minimum number of samples required to split an internal node, a smaller value would enable the model to fit the data more *finely*.
2. The best models seem to be robust to changes in the `n_estimators` because the top 3 models all have different `n_estimators`. The `n_estimators` determines the number of trees in the forest, and it was interesting to note that the model fits didn't change considerably when the parameter was changed. This results is corroborated by (Cutler et al., 2012) and it can be speculated that this robustness arises as a result of *bagging* performed to reach a consensus.
3. The best models have the `max_features` parameter set to `sqrt` as opposed to `log2`. Taking into consideration the number of features in our dataset, the value returned by `sqrt` is higher than that returned by `log2`. This also indicates that when information from larger number of features (time points, in our case) are used, it results in better F1 score and accuracy.

3.3 K-Nearest Neighbors

The top 5 model, tuned with parameters obtained using Grid Search are as follows:

Name	Neighbors	Weights	F1	Accuracy
KNN	10	distance	0.84	0.81
KNN_SS	10	distance	0.84	0.81
KNN	15	distance	0.82	0.79
KNN_ND	3	distance	0.81	0.78
KNN_SS	5	distance	0.81	0.79

Table 3: Top 5 performances of the KNN Classifier model, sorted based on validation F1 scores, for different parameter values.

Interpretations based on the best model parameters:

1. The best models have the `n_neighbors` parameter value of 10. When this is increased or decreased, there is a small dip in the F1 score and accuracy. This implies that only the closest 10 neighbors are relevant to the datapoint of interest, thereby giving us an intuition about the distribution of the datapoints in our dataset.
2. All the top 5 models used the `weights` parameter of `distance` as opposed to `uniform`. This highlights that the relative distance between the datapoints is more informative than when compared to using a uniform weight distribution, when making a prediction.

3.4 Support Vector Machines

The top 5 model, tuned with parameters obtained using Grid Search are as follows:

Name	C	Decision Function	Degree	γ	Kernel	F1	Accuracy
SVC_SS	10	ovo	3	auto	poly	0.84	0.85
SVC_SS	10	ovr	3	auto	poly	0.84	0.85
SVC	10	ovo	3	scale	poly	0.82	0.83
SVC_SS	10	ovo	3	scale	poly	0.82	0.83
SVC_SS	10	ovr	3	scale	poly	0.82	0.83

Table 4: Top 5 performances of the SVM Classifier model, sorted based on validation F1 scores, for different parameter values.

Interpretations based on the best model parameters:

1. The value of 10 for the `C`, inverse regularization parameter seemed to give raise to best predictions. The next best values of `C` reported by grid search are 50, followed by 0.5 and 0.1. This implies that the sweet spot for the regularization parameter is 10, and higher values would result in higher variance and lower values would result in higher bias.
2. All the 5 top models used a `degree` of 3. This in conjunction to the `kernel` being set to `poly` gives us an intuition that the best mapping from the actigraph space to the classification is a non-linear, polynomial mapping of degree 3.
3. The best models were robust to the `decision_function_shape` used. This is intuitive as all the models were trained on a binary-class classification task.
4. The parameter `gamma` is the inverse of the radius of influence of a single training sample. When large values of `gamma` is used, the radius of influence becomes small and results in overfitting. Similarly, when small values of `gamma` is used, the radius of influence becomes large and results in underfitting. The value of `auto` as opposed to `scale` best fits the dataset and implies that it is not essential to make `gamma` dependent on the variance of the dataset.

3.5 Multi-layered Perceptrons

The top 5 model, tuned with parameters obtained using Grid Search are as follows:

Name	Activation	α	# Nodes	LR Scheduler	F1	Accuracy
MLP_ND	relu	0.0001	(50, 10)	invscaling	0.81	0.84
MLP_SS_ND	relu	0.0001	(50, 10)	adaptive	0.81	0.84
MLP_SS_ND	relu	0.0001	(50, 10)	invscaling	0.81	0.84
MLP_SS_ND	relu	0.0001	(50, 10)	constant	0.81	0.84
MLP_ND	relu	0.0001	(50, 10)	adaptive	0.81	0.84

Table 5: Top 5 performances of the MLP Classifier model, sorted based on validation F1 scores, for different parameter values.

Interpretations based on the best model parameters:

1. The best models all utilized the `relu` activation function as opposed to `tanh`. This in in conjunction with the recent trends in deep learning (Fukushima, 1980), (Nair and Hinton, 2010), (Agarap, 2018).
2. Lower value of `alpha`, coefficient of regularization, has been opted by all the best models.
3. All the best models use a `hidden_layer_sizes` of (50, 10). This shows that the model with two layers performs better than when compared to models that have a single hidden layer. This potentially ties to the models' ability to extract more information from the dataset in the presence of additional layers.

4. The models seem to be robust to the choice of `learning_rate` scheduler used. However, from the results, we see that `invscaling` and `adaptive` are better represented than `constant`, indicating that dynamic adjustments to learning rates is more advantageous than static learning rates.

3.6 Model Selection

The top 15 models and parameter combinations presented in Figure 4 were subsequently used to address all the proposed hypothesis. Across all models, when PCA was applied on the data prior to model fitting, the performance dropped. Hence, PCA based dimensionality reduction was not used in all subsequent analysis.

4 Results Visualization

4.1 Hypothesis 1

The model performance of Random Forest Classifiers and Logistic Regression Classifiers was obtained. As observed in the previous section, the best model performance are as follows:

- Logistic Regression: F1 Score: 0.77; Accuracy: 0.83
- Random Forest: F1 Score: 0.89; Accuracy: 0.90

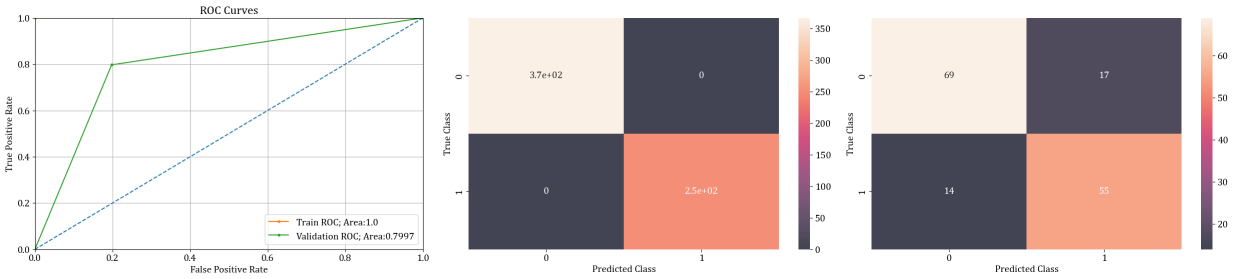

Figure 5: ROC Curve and Confusion Matrix (train, validation) for the best model; Hypothesis 1

As we can see above, the best Random Forest model provides a really good fit to the dataset and is able to perform better, both in terms of accuracy and F1 score, than when compared to Logistic Regression.

4.2 Hypothesis 2

Since the data is arranged in a day-wise format across all participants, the age and gender information for all participants was tiled and stacked with the actigraph dataset. This dataset comprising actigraph information, age and gender data was used for analysis in this section.

When the age and gender data was included, both the F1 score and accuracy increased slightly. The best model results are: F1 Score: 0.925926 and Accuracy: 0.935484. The ROC Curve and Confusion Matrix of the best model is shown in Figure 6. From the confusion matrix we can see that the model is able to correct some of the wrongly classified data points when the age and gender information is added.

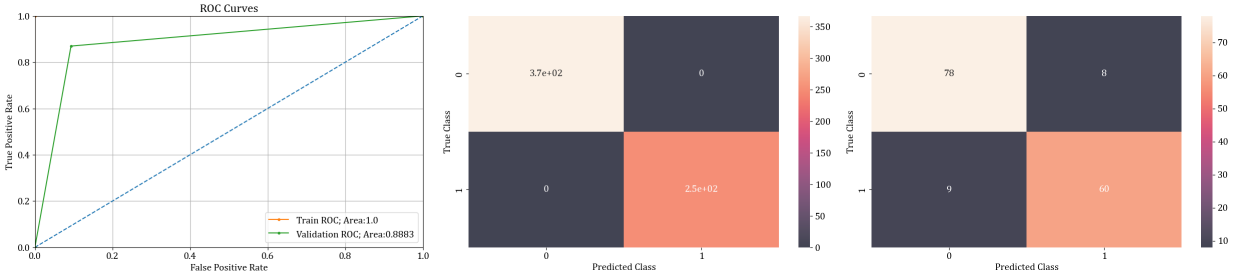

Figure 6: ROC Curve and Confusion Matrix (train, validation) for the best model; Hypothesis 2

4.3 Hypothesis 3

The same dataset generated for the last hypothesis was used here. The best model results are: F1 Score: 0.925926 and Accuracy: 0.935484. The ROC Curve and Confusion Matrix of the best model is shown in Figure 6. The accuracies and F1 scores for males and females obtained from the model that best fit the dataset is as follows:

- Male: Accuracy: 0.953846; F1 Score: 0.958904
- Female: Accuracy: 0.831325; F1: 0.758621

4.4 Hypothesis 4

A similar approach as in Hypothesis 2 was taken to generate a dataset that had information about the level of depression in a patient. Both the MADRS Scores 1 and 2 were included in the dataset.
The accuracies and F1 scores for mild and severe depression obtained from the model that best fit the dataset (same model as the previous subsection) is as follows:

- Mild Depression: Accuracy: 0.972727; F1: 0.938776
- Severe Depression: Accuracy: 0.822222; F1: 0.902439

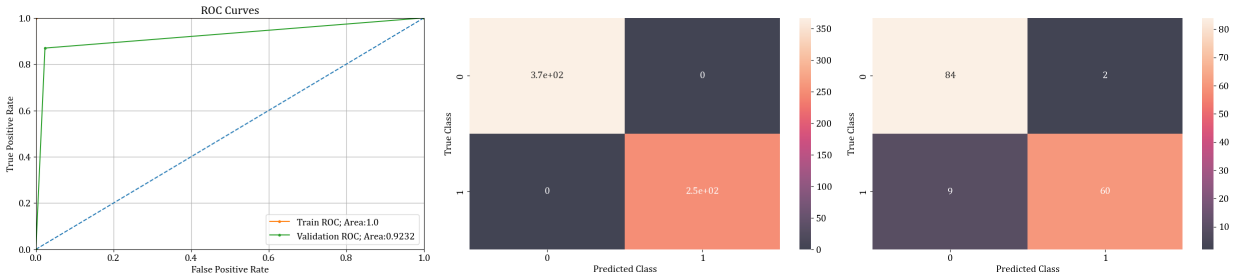

Figure 7: ROC Curve and Confusion Matrix (train, validation) for RF model.

5 Discussion

This study aimed to determine which classification model could most accurately predict depression based on motor activity recorded from an Actigraph wearable wristband. Our study used motor Actigraph data from (Garcia-Ceja et al., 2018).

In our analysis, we used 5 different classes of models, namely Logistic Regression, Random Forest, K-Nearest Neighbors Classifier, Support Vector Machine Classifier, and Multi-Layer Perceptron Classifier. We sought to determine which model parameters would enable the best predictor of depression from motor activity. We utilized a dataset comprising 773 data points, with 340 belonging to the depressed condition and the remaining samples serving as controls.

As the number of features (activity per minute) in our data was 1440, representing a high-dimensional data space, we attempted to use Principal Component Analysis (PCA) to reduce the dimensionality of our dataset and identify the most relevant features. However, we found that this approach did not significantly improve our model's performance and did not use PCA based dimensionality for our subsequent analyses. This indicates that specific motor activity features are critical for distinguishing between individuals with depression and those without. Jointly considering the `kernel` results from `SVC` and the results from `PCA`, we can see that the mapping between the actigraph space and the output space is non-linear. Which could also explain why PCA, a linear dimensionality reduction technique did not result in better classification accuracies.

The best parameters for each model were identified using grid search, combined with a 5-fold cross validation on the training sample. The models that best differentiated between the control and condition groups are the Random Forest models, Support Vector Machine Classifier models and K-Nearest Neighbors Classifier models, in the same order.

We found that the random forest model could accurately classify the samples' condition status with a high F1 score and accuracy. The superior performance of the random forest model can be attributed to its ability to handle high-dimensional data by constructing a large number of decision trees and the use of bagging to combine their results to make the final classification decision. This approach is advantageous when dealing with complex data structures where traditional regression models may not be sufficient. Our findings support the use of RF models as a powerful tool for classification tasks in mental health research when the dataset is small and has a large number of features.

We also developed a logistic regression model to predict depression based on motor activity. We investigated the impact of regularization and solver algorithms on the model's predictive performance. Our findings indicate that logistic regression was robust to both the coefficient and type of regularization, i.e., `L1` and `L2` regularization, but `L2` utilizes a larger regularization coefficient than `L1` to achieve the same level of performance. We compared two solver algorithms for logistic regression, `liblinear` and `saga`. Our results indicate that on our small dataset `liblinear` performs better than `saga`. However, when the data was not normalized or standard scaling was not used, `saga` performed better than `liblinear`. This could be attributed to `liblinear`'s tendency to converge at a non-stationary point, leading to suboptimal model performance. Our findings suggest that while logistic regression is a robust model for predicting depression based on motor activity, it is outperformed by the RF model.

We further investigated the impact of adding age and sex data to our predictive models for depression classification based on motor activity. Our findings indicate that including age and sex data improved the model's classification performance. Additionally, we observed that our models had higher accuracies and F1 scores for males than females. This suggests that there may be gender-based differences in the impact of motor activity on depression. We also investigated the model performances in accurately predicting mild and severe depression. We found that when control samples were included, the accuracy and F1 score of mild depression cases were higher than severe depression cases. This suggests that including control samples may provide additional information that could aid in better differentiating between mild and severe depression cases based on motor activity.

While our study provides valuable insights into using RF models for depression classification, several limitations must be acknowledged. Our dataset was relatively small, and future studies with larger sample sizes are required to confirm our findings. Furthermore, there are limitations to using motor activity as the sole feature for depression classification. Our dataset does not account for the individual variability in movement and fails to account for other symptoms of depression, such as one's mood or cognitive state. Additionally, we did not explore the potential impact of confounding variables on our results, such as medication use or comorbid conditions. Future studies should address these limitations and further refine our understanding of the role of motor activity in depression.

References

Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Aminifar, A., Rabbi, F., Pun, V. K. I., and Lamo, Y. (2021). Monitoring motor activity data for detecting patients' depression using data augmentation and privacy-preserving distributed learning. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2163–2169. IEEE.

Cutler, A., Cutler, D. R., and Stevens, J. R. (2012). Random forests. *Ensemble machine learning: Methods and applications*, pages 157–175.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear: A library for large linear classification. *the Journal of machine Learning research*, 9:1871–1874.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202.

Garcia-Ceja, E., Riegler, M., Jakobsen, P., Tørresen, J., Nordgreen, T., Oedegaard, K. J., and Fasmer, O. B. (2018). Depresjon: a motor activity database of depression episodes in unipolar and bipolar patients. In *Proceedings of the 9th ACM multimedia systems conference*, pages 472–477.

Marrie, R., Walld, R., Bolton, J., Sareen, J., Walker, J., Patten, S., Singer, A., Lix, L., Hitchon, C., El-Gabalawy, R., et al. (2019). Rising incidence of psychiatric disorders before diagnosis of immune-mediated inflammatory disease. *Epidemiology and psychiatric sciences*, 28(3):333–342.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.