UPDATED RESULTS

In the following model I used a new abstraction method I implemented that is based on labeling time intervals in each entity, frequency band and channel based on their frequency power. These results are based on counting the labels for each subject as features and still without using KarmaLego patterns.

Table 1: model results for different classification models on above **median change in HDRS21 score (H7)**. Features from **Frequency Power** abstractions, subject records are full records. Data imbalance is 51% above median and 49 % below median (median was calculated on both H7 and H1). Train test ratio is 80/20.

Each model results were calculated 10 times with different train test split and the results above are mean $\pm$ std. Data size contains 82 records.

Model	**Accuracy**	**F1**	**Recall**	**Precision**
XGBoost	0.61 ± 0.12	0.60 ± 0.14	0.61 ± 0.12	0.61 ± 0.17
RandomForest	0.58 ± 0.13	0.55 ± 0.15	0.57 ± 0.13	0.57 ± 0.18
Catboost	**0.61 ± 0.13**	**0.59 ± 0.15**	**0.61 ± 0.14**	**0.61 ± 0.14**
TPF	0.56 ± 0.12	0.55 ± 0.14	0.56 ± 0.13	0.56 ± 0.16
TPF_top5	0.54 ± 0.07	0.50 ± 0.08	0.53 ± 0.07	0.53 ± 0.09
TPF_top7	0.60 ± 0.10	0.57 ± 0.13	0.59 ± 0.11	0.59 ± 0.15
TPF_top9	0.58 ± 0.12	0.54 ± 0.14	0.57 ± 0.12	0.58 ± 0.18
TPF_top11	**0.63 ± 0.09**	**0.60 ± 0.10**	**0.62 ± 0.09**	**0.68 ± 0.12**
TPF_top13	0.59 ± 0.10	0.56 ± 0.12	0.58 ± 0.11	0.60 ± 0.12

Table 2: model results for different classification models on above **median change in HDRS21 score (H1)**. Features from **Frequency Power** abstractions, subject records are full records. Data imbalance is 41% above median and 59 % below median (median was calculated on both H7 and H1). Train test ration is 80/20.

Each model results were calculated 10 times with different train test split and the results above are mean $\pm$ std. Data size contains 57 records.

Model	**Accuracy**	**F1**	**Recall**	**Precision**
XGBoost	0.50 ± 0.08	0.45 ± 0.09	0.47 ± 0.08	0.46 ± 0.12
RandomForest	0.53 ± 0.14	0.43 ± 0.16	0.48 ± 0.14	0.43 ± 0.22
Catboost	0.56 ± 0.11	0.51 ± 0.14	0.53 ± 0.12	0.53 ± 0.19
TPF	0.44 ± 0.11	0.42 ± 0.12	0.43 ± 0.12	0.41 ± 0.14
TPF_top5	0.57 ± 0.15	0.55 ± 0.15	0.56 ± 0.15	0.57 ± 0.16
TPF_top7	**0.58 ± 0.11**	**0.55 ± 0.11**	**0.56 ± 0.11**	**0.58 ± 0.14**

Table 3: model results for different classification models on above **median change in HDRS21 score (H7)**. Features from **Frequency Power** abstractions, subject records are divided to 1-minute records assuring no data leakage (same subject in train and test). Data imbalance is 51% above median and 49 % below median (median was calculated on both H7 and H1). The train test ration is 80/20.

Each model results were calculated 10 times with different train test split and the results above are mean $\pm$ std. Data size contains 309 records.

Model	**Accuracy**	**F1**	**Recall**	**Precision**
XGBoost	0.51 ± 0.09	0.50 ± 0.09	0.51 ± 0.09	0.51 ± 0.10
RandomForest	0.49 ± 0.07	0.48 ± 0.07	0.49 ± 0.07	0.49 ± 0.08
Catboost	0.51 ± 0.06	0.50 ± 0.06	0.51 ± 0.06	0.51 ± 0.06
TPF	0.47 ± 0.09	0.46 ± 0.10	0.47 ± 0.09	0.47 ± 0.10
TPF_top5	0.62 ± 0.05	0.61 ± 0.05	0.62 ± 0.05	0.63 ± 0.06
TPF_top9	0.54 ± 0.09	0.53 ± 0.08	0.54 ± 0.09	0.54 ± 0.09
TPF_top7	0.57 ± 0.07	0.57 ± 0.07	0.57 ± 0.07	0.57 ± 0.08
TPF_top3	**0.63 ± 0.06**	**0.62 ± 0.06**	**0.63 ± 0.06**	**0.64 ± 0.07**
TPF_top1	0.60 ± 0.06	0.59 ± 0.06	0.60 ± 0.06	0.61 ± 0.07

Table 4: model results for different classification models on above **median change in HDRS21 score (H7)**. Features from **Frequency Power** abstractions, subject records are divided into **1 minute with 30 seconds overlap** records assuring no data leakage (same subject in train and test). Data imbalance is 51% above median and 49 % below median (median was calculated on both H7 and H1). The train test ration is 80/20.

Each model results were calculated 10 times with different train test split and the results above are mean $\pm$ std. Data size contains 582 records.

*You can notice the result didn't improve with the overlap

Model	Accuracy	F1	Recall	Precision
XGBoost	0.46 ± 0.07	0.45 ± 0.07	0.46 ± 0.07	0.46 ± 0.07
RandomForest	0.51 ± 0.08	0.50 ± 0.08	0.51 ± 0.08	0.51 ± 0.08
Catboost	0.50 ± 0.07	0.49 ± 0.07	0.50 ± 0.07	0.50 ± 0.07
TPF	0.41 ± 0.09	0.40 ± 0.10	0.41 ± 0.10	0.41 ± 0.10
TPF_top3	0.61 ± 0.09	0.60 ± 0.10	0.61 ± 0.09	0.61 ± 0.10
TPF_top7	**0.62 ± 0.09**	**0.61 ± 0.10**	**0.62 ± 0.09**	**0.62 ± 0.10**
TPF_top11	0.59 ± 0.10	0.59 ± 0.10	0.59 ± 0.10	0.59 ± 0.11

Table 5: model results for different classification models on above **median change in HDRS21 score (H1)**. Features from **Frequency Power** abstractions, subject records are divided into **1 minute with 30 seconds overlap** records assuring no data leakage (same subject in train and test). Data imbalance is 41% above median and 59 % below median (median was calculated on both H7 and H1). The train test ration is 80/20.

Each model results were calculated 10 times with different train test split and the results above are mean $\pm$ std. Data size contains 464 records.

Model	Accuracy	F1	Recall	Precision
XGBoost	0.60 ± 0.07	0.57 ± 0.08	0.58 ± 0.07	0.58 ± 0.08
RandomForest	**0.62 ± 0.10**	**0.59 ± 0.11**	**0.60 ± 0.10**	**0.60 ± 0.12**
Catboost	0.61 ± 0.08	0.57 ± 0.09	0.58 ± 0.08	0.59 ± 0.10
TPF	0.54 ± 0.13	0.51 ± 0.13	0.46 ± 0.13	0.46 ± 0.14
TPF_top3	**0.59 ± 0.04**	**0.57 ± 0.05**	**0.57 ± 0.05**	**0.57 ± 0.05**
TPF_top5	0.56 ± 0.05	0.54 ± 0.06	0.54 ± 0.06	0.54 ± 0.06
TPF_top9	0.56 ± 0.07	0.54 ± 0.07	0.55 ± 0.07	0.55 ± 0.07

Important thing to notice here is that even though the data is imbalanced the results are above 50% in all metrics and not only on accuracy.

Table 6: model results for different classification models on above **median change in HDRS21 score (H1)**. Features from **Frequency Power** abstractions, subject records are divided into **1 minute with 30 seconds overlap** records assuring no data leakage (same subject in train and test). Data imbalance is 51% above median and 49 % below median (median was calculated H1 only). Here I removed from training the data in percentage of between 35-65 % of the HDRS21_change distribution so the model will define better between above median and below median.

Each model results were calculated 10 times with different train test split and the results above are mean $\pm$ std. Data size contains 262 records (due to the change I mentioned above train test ratio was 60/40).

Model	Accuracy	F1	Recall	Precision
Catboost	**0.60 ± 0.09**	**0.58 ± 0.09**	**0.59 ± 0.09**	**0.62 ± 0.13**
XGBoost	0.55 ± 0.10	0.52 ± 0.10	0.54 ± 0.09	0.56 ± 0.13
RandomForest	**0.59 ± 0.09**	**0.55 ± 0.11**	**0.58 ± 0.09**	**0.64 ± 0.14**
TPF	0.54 ± 0.13	0.52 ± 0.13	0.50 ± 0.15	0.50 ± 0.15
TPF_top3	**0.62 ± 0.11**	**0.59 ± 0.13**	**0.62 ± 0.11**	**0.63 ± 0.15**
TPF_top5	0.60 ± 0.12	0.56 ± 0.15	0.59 ± 0.12	0.59 ± 0.18
TPF_top7	0.60 ± 0.13	0.57 ± 0.16	0.60 ± 0.13	0.62 ± 0.19
TPF_top11	**0.62 ± 0.13**	**0.58 ± 0.15**	**0.62 ± 0.13**	**0.66 ± 0.16**

Conclusions:

We can see that almost all models achieved results higher than 50%, specifically the models who worked best were TPF top k features which might imply on the advantage of predicting by similarity to specific former subjects of whom we already have results. Another important thing to notice is that in the H7 model that has relatively big dataset (82 subjects), the division to parts of records made the model worse rather than better. In H1 with a relatively small dataset (57 subjects) the division improved the model. This implies the importance of long records in predicting the class of individuals (in this case above median change in HDRS21 score).

Important note is that these results are based on the label count from abstraction and not yet with the temporal patterns between labels. So hopefully soon there will be even better results.