Automated diagnosis of schizophrenia from interlocutors’ coordination

Dominik Klepl

Supervised by Riccardo Fusaroli

Bachelor Thesis, Cognitive science, Aarhus University

2018-12-15

**Contents**

[1 Abstract 4](#_Toc532819600)

[2 Introduction 5](#_Toc532819601)

[2.1 Definition of schizophrenia and diagnosis 6](#_Toc532819602)

[2.2 Alternative markers as proxy for symptoms 6](#_Toc532819603)

[2.3 Social impairment in schizophrenia 7](#_Toc532819604)

[2.3.1 Social impairment as emerging phenomenon 8](#_Toc532819605)

[2.4 Machine learning and interpretability 9](#_Toc532819606)

[3 Methods 11](#_Toc532819607)

[3.1 Data 11](#_Toc532819608)

[3.2 Preprocessing 11](#_Toc532819609)

[3.3 Feature Extraction 12](#_Toc532819610)

[3.3.1 Recurrence quantification analysis 12](#_Toc532819611)

[3.4 Data partitioning 13](#_Toc532819612)

[3.5 Feature Selection 13](#_Toc532819613)

[3.5.1 Filtering 14](#_Toc532819614)

[3.5.2 Method 1 Boruta 14](#_Toc532819615)

[3.5.3 Method 2 Filter correlation 14](#_Toc532819616)

[3.5.4 Method 3 Featexp 14](#_Toc532819617)

[3.6 Model training 15](#_Toc532819618)

[3.7 Evaluation 15](#_Toc532819619)

[3.8 Model interpretation 16](#_Toc532819620)

[4 Results 18](#_Toc532819621)

[4.1 Preprocessing tests 18](#_Toc532819622)

[4.2 Feature selection 18](#_Toc532819623)

[4.2.1 Participant feature selections 18](#_Toc532819624)

[4.2.2 Coordination feature selections 18](#_Toc532819625)

[4.3 Evaluation of the feature selections 19](#_Toc532819626)

[4.4 Participant vs. coordination models on validation data 19](#_Toc532819627)

[4.5 Model interpretation 19](#_Toc532819628)

[5 Discussion 20](#_Toc532819629)

[5.1 Discussion of the results 20](#_Toc532819630)

[5.2 Interviewer bias 21](#_Toc532819631)

[5.3 Retraining the models 22](#_Toc532819632)

[6 Conclusion 23](#_Toc532819633)

[7 References 24](#_Toc532819634)

[8 Appendix 28](#_Toc532819635)

# Abstract

# Introduction

Schizophrenia is defined by collection of various symptoms and during a typical diagnosis a trained professional is assessing whether and to what extend the subject displays these symptoms. This system of diagnosing is often criticized and doubted even by mental health clinicians who follow these symptom-based manuals on daily basis (Tzur Bitan et al., 2018). One of the main arguments is the proneness of this system to fail to diagnose or misdiagnose the patient, because the symptoms often overlap between multiple disorders (Fried & Nesse, 2015), and fear of administering drugs that might hurt the patient. This is one of the reasons for an extensive research on alternative markers of mental disorders such as genetics (Delude, 2015) that could be objectively measured. Impairment of social function is linked to schizophrenia (Brunet-Gouet & Decety, 2006) and can be observed for example in the speech of the patients (Parola, Simonsen, Bliksted, & Fusaroli, 2018). Another consequence of the social function deficit is disrupted non-verbal communication. Traditionally it was assumed that communication and implicitly social function can be explained by looking at one interlocutor only (Deacon & Poeppel, 1997; Gallese & Lakoff, 2005). Recent findings indicate, however, that social function is not property of an individual but a phenomenon emerging from social interactions (Fusaroli, Rączaszek-Leonardi, & Tylén, 2014). In other words, to understand social function one has to look at all parties involved in the communication and how they coordinate with each other. This thesis follows this suggestion by hypothesizing that although hand gestures of the patient alone are informative of the diagnosis, the coordination of gestures between the patient and interviewer during a social interaction is better predictor. To test the hypothesis, this thesis proposes using hand gestures of patient and interviewer and their coordination recorded during a clinical interview to build an automated system for diagnosing schizophrenia. Furthermore, instead of building models that would simply output a probability of the diagnosis, the thesis aims to produce transparent machine-learning models whose decisions can be interpreted and justified. Apart from supporting the hypothesis the machine-learning based models reported in this thesis might prove useful in real-life application. They could be utilized by the diagnosticians to improve and support their decision making.

The thesis first summarizes the ICD-10 definition (World Health Organization, 2016) of schizophrenia. It then presents the current research on other markers of the disorder along with attempts on utilizing these markers for producing AI diagnosticians of mental disorders. It then presents the evidence of schizophrenics gesturing differently from healthy population and introduces the emerging synergy theory which leads to the thesis’ hypothesis described above. This section also presents the justification for the features used for training the models. It then makes the case for using machine learning models in addition to statistical models and the emerging need for interpretable models. It then continues with presenting the methods used for data preprocessing, training the predictive models, their evaluation and interpretation. The results are then discussed together with limitations of the approach and suggested solutions.

## Definition of schizophrenia and diagnosis

Schizophrenia or schizotypal disorder is characteristic by inappropriate and/or blunt affects and distorted thinking and perception (*Schizophrenia Fact sheet N°397*, 2015). There is more than 23 million people diagnosed with schizophrenia in the world. The disorder begins to manifest in age between 18-25 years for men and 25-35 years for women and it is more common among males who also tend to exhibit more severe symptoms. The symptoms are commonly divided in two categories; positive and negative. Among the positive symptoms belong hallucinations and delusions. Some of the common types of positive as reported by the patients are voices arguing about the patient and referring to her in 3rd person or commenting the patient’s thoughts or behavior (Schneider, 1959). On the other hand, the negative symptoms are apathy, lack of emotions or inappropriate social skills.

To this day there is no physical or lab test that is certain to diagnose schizophrenia (American Psychiatric Association, 2013). The diagnosis is rather issued by a professional diagnostician who often speaks to the patient with the target to identify the symptoms listed either in ICD-10 (World Health Organization, 2016) or DSM-5 manuals for diagnosis. One of the problems of this approach is the subjective nature of the evaluation of the strength of the symptoms. Many of the positive symptoms can only be reported by the patient. There are, however, attempts for developing quantifiable and accurate tests, not only of schizophrenia but other mental disorders whose diagnosis systems suffer from similar problems, as described in the following section.

## Alternative markers as proxy for symptoms

As mentioned above, the current way of diagnosis is not optimal which is why there are several projects to develop tests that do not rely on self-reports that could be used in addition to the current method. These tests are not explicitly intended to diagnose given disorder. They rather target specific symptoms such as avolition because symptoms are less complex constructs than diagnoses defined as complex mixtures of symptoms. There are attempts to develop blood tests for detecting schizophrenia (Tsuang et al., 2005) that measure activity of certain genes. A meta-analysis suggests it is possible to predict schizophrenia from an intelligence test up to three years before the onset (Woodberry, Giuliano, & Seidman, 2008). There is also an ongoing project leveraging the recent development in smart wearable technology to monitor bipolar disorder patients which allows the psychiatrist an access to data that could never be collected during a clinical observation and can lead to better targeted treatment (Puiatti, Mudda, Giordano, & Mayora, 2011). However, there is no project, to the author’s knowledge, that is using hand gestures and interlocutor coordination to predict schizophrenia. Use of hands as means of nonverbal communication is believed to be part of social function. Therefore, the research on social function impairment in schizophrenia is covered in the next section.

## Social impairment in schizophrenia

As mentioned above, impairment of social function is often linked to schizophrenia which often lead to social isolation in which the patients often communicate only with close family. It was suggested that this deficit is connected to deficits in social cognition, specifically to theory of mind emotion recognition abilities (Bora & Pantelis, 2016). Tasks such as understanding emotional expressions or understanding complex language constructs such as sarcasm pose problems for schizophrenia patients. Another expression of impaired social function can be observed in the way schizophrenia patients speak (Parola et al., 2018). They exhibit acoustic patterns such as flatter intonation, longer pauses and higher pitch. One can observe distinctively different behavior of schizophrenics during social interactions (Lavelle, Healey, & McCabe, 2014). The patients tend to speak less than the controls. Furthermore, the patients’ partners experienced poorer rapport. Differences can be observed also in communication disfluencies in dialogs such as unfilled pauses between utterances that are more pervasive in dialogs with a patient (Howes, Lavelle, Healey, Hough, & McCabe, 2017). On the other hand, patients use less filled pauses that controls tend to fill with ‘um’ or ‘ehm’. These results suggest that it is more difficult to achieve smooth coordination of turn taking. Another expression of impaired social function is the non-verbal communication. It was shown that patients coordinate their hand movements with speech less than controls (Lavelle, Howes, Healey, & McCabe, 2013). This study also showed that the partners talking to the patients appear to adopt this decreased hands-speech coordination but to a less extent. It seems that non-verbal communication is disrupted in patients with schizophrenia (Lavelle, Healey, & McCabe, 2012). The patients speak and gesture less while speaking and also nodded less as listeners while their partners showed an opposite pattern which could be understood as an attempt to compensate the deficit of non-verbal communication during the interaction. These findings motivated the extracted features for training the models. There is generally a pattern of less gesturing in patients which can be quantified by descriptive statistics features such as mean, median but also interquartile range or median absolute deviation because those would capture the difference even in case of nonstationary data. There is also often the negative relationship between gestures of patient and patient’s partner which can be quantified by taking difference of the descriptive statistics calculated from patient’s gestures and partner’s gestures.

### Social impairment as emerging phenomenon

Many studies report some degree of impaired coordination and compensation for the deficits of social function during an interaction (Howes et al., 2017; Lavelle et al., 2012; Lavelle et al., 2013). Traditionally it was assumed that social function is a property of an individual (Deacon & Poeppel, 1997; Gallese & Lakoff, 2005). But the findings reported above and other seem to indicate that social function is rather something establishing in the course of social interaction. This view is introduced in interpersonal synergy theory (Fusaroli, Rączaszek-Leonardi, et al., 2014). This theory attributes the social function not to the cognitive systems of the interlocutors but rather to a gradually emerging interpersonal system between them. From this theory follows the hypothesis of this thesis:

Coordination is what underlies the social function therefore it contains more information than individual signal which is only one component of an interpersonal system.

The thesis operationalizes this hypothesis by saying that model trained on features relating to the coordination between the interlocutors yields higher predictive power towards the diagnosis of schizophrenia than a model trained on features extracted from the gesture of the patient. Predictive power is defined as error when predicting data previously unseen by the model.

The interpersonal synergy theory also motivated another category of features to extract from the gesture signals as it predicts the interpersonal system to be dynamic and non-linear. The cross recurrence quantification analysis (RQA) allows to extract features describing linear and non-linear relationships between 2 signals (Fusaroli, Konvalinka, & Wallot, 2014). This method can be applied to the patient’s signal as well to quantify the repetitive patterns. It was previously showed that RQA provides useful features for exploring voice dynamics of schizophrenics (Fusaroli, Weed, Simonsen, & Bliksted, 2013) but has not been previously applied to gesture signals therefore we there are no theory-based predictions what RQA features might show.

## Machine learning and interpretability

In addition to using statistical models for binomial classification such as logistic regression this thesis is also using machine learning (ML) algorithms to build the classifiers. To cover high range of possible patterns in the data and different types of algorithms 5 types of models were selected. First is the L2 penalized logistic regression. The advantage of this model is its simplicity and fast training while also allowing unimportant features to be dropped completely from the model. The simplicity is however also its main weakness as it can model only linear relationships. Second is the naive Bayes classifier. This model is also very quick to fit but has some heavy assumptions for numerical features such as normality and independence. In case of violation of the assumptions the model can overestimate importance of features leading to overfitting. But at the same time, it has proved to perform better than more complex algorithms even with assumptions violated. Third is the random forest which is more complex but still allows quick fitting, mainly because of parallel computing support (Nielsen, 2016). The weakness lies in increasing overfitting with the depth of the trees and its preference for features with high number of unique values. Next model is extreme gradient boosting machine (XGBoost) which is the most state-of-the-art model in our selection. We are using the random forest version of the XGBoost which is able to decrease overfitting and variance by regularizing the trees (both L1 and L2) and fitting only shallow decision trees (Nielsen, 2016). The fourth model is model averaged neural networks (avNNet), using averaged predictions of the same neural network fitted with several random seeds (Ripley, 2007). Strength of avNNet is that while it explores the different shapes of patterns that could be fit to the data it also smooths the line of fit enough to prevent overfitting. The weakness is mainly the lower speed of fit.

Most of these models are so-called black boxes, in other words we do not know the decision processes happening inside of the model, only the output. In recent years, the importance of producing interpretable machine learning models has been increasingly pointed out (Doshi-Velez & Kim, 2017; Gunning, 2017; Ribeiro, Singh, & Guestrin, 2016). If ML should be used in scientific research regularly, model interpretability is crucial for the research to produce any insights. Second, models whose predictions are justifiable are necessary for ML applications outside of academia. E.g. for schizophrenia classifier to be deployed for daily use by psychiatrists the model has to be able to explain the user why she should consider adjusting the original decision. The former can be reached using global-level explainers that explore the general patterns learned from the data. On the other hand, explainers at prediction-level that can breakdown the final prediction into contributions of the individual features aim for the latter.

The thesis is utilizing both categories of explainers, however as we do not have specific enough predictions, the results are not commented in depth but rather intended to show that both goals of ML interpretability can be reached quite well with current methods. More specifically the thesis is using feature permutation and ALE plots to showcase the global-level methods, and game-theory based SHAP decomposition as an example of prediction-level methods.

# Methods

## Data

The data was originally collected for Simonsen et al. (2018). The dataset comprises of 41 patients diagnosed with ICD-10 DCR diagnosis of schizophrenia or schizoaffective disorder and 43 healthy controls. The patients were selected through the Psychiatric Centre of the National Hospital of the Faroe Islands. The data was collected during a clinical interview, Schedules for Clinical Assessment in Neuropsychiatry (SCAN). All the interviews were led by the same psychologist. Movement of participant’s and interviewer’s hands were recorded with an actigraph with sampling rate of 100 Hz attached to the wrists during the whole session. Only data from the dominant hand was used as the study was not concerned with coordination within one person. 76 participants are right-handed, 8 participants and the interviewer are left-handed. The average age of the participants was 38 (sd = 10.3) with range between 18 and 55 years. The interview lasted 58.5 minutes on average (sd = 39.3) with the shortest having 17.6 minutes and the longest having 163.68 minutes.

## Preprocessing

To be able to extract features from the actigraph recordings, several preprocessing steps were applied. The actigraph captures movement as acceleration on 3-axis. These 3 axes are, however, meaningless as they change with every movement. Therefore, an Euclidian distance of the 3 points were taken. Then the signals had to be aligned in time using an audio recording of the interview because the actigraph recording was often started quite long before the interview actually begun. This was possible because the beginning of the interview was marked by the interviewer clapping three times rhythmically which was clearly visible both in audio signal and actigraph signal. A combination of automated clap detection algorithms and manual validation were used to find the claps in both types of signals to ensure the claps were identified correctly in all signals. The claps in actigraph signals were found using a regular-expressions based algorithm that searched for a typical shape of the claps in the data that represented the claps. The claps had shape of 3 prominent spikes with similar distance between their peaks. The top of the spike was defined as point whose neighboring point to the left has less than 80 % of its value and 5 points from the right have lower value than the top of the spike. The algorithm is the adjusted version of code from R package pracma (Borchers, 2018). This approach, however, does not perform well in locating the claps in the audio signal. For audio signals the algorithm implemented in the R package (Andrey, 2018) was used. Although the algorithm is originally intended for syllable separation in voice recording it was found to perform well in finding claps. The algorithm locates claps by looking for sudden bursts of acoustic energy. Its output is not only the location of the bursts but also the interval between the bursts. Therefore, the claps were defined as three large bursts with similar interburst intervals. All identified automatically identified claps were manually inspected to avoid any misclassification. These checks were performed by plotting the signal with the claps tagged as in figure X and manually confirming the position of the first clap. After the locations of all claps were identified, the actigraph signals were trimmed so that the first data point was immediately after the third clap occurred. The aligned signals were then tagged by interlocutor speaking. This was again enabled by the audio recordings of all the interviews. Using an automatized diarization system the audio recordings were split into utterances and tagged by speaker. This output was then used to tag the actigraph data. Those parts of the data where nobody was speaking were excluded as it can be assumed that no gesticulation occurred during that time. Some sequences were identified, totaling to 67 minutes, when the actigraph was not recording and therefore these parts of the signals had to be removed as well. Finally, the signals were split into 10 seconds long samples. Because the class imbalance of the target variable was high (70% samples from schizophrenics) the data was downsampled.

## Feature Extraction

21 statistical features and 8 recurrence quantification analysis (RQA) features were extracted *(see Recurrence quantification analysis)* and included in the set of features using only the participant’s signal (further referred to as PF). In the set of features using coordination between the participant’s and interviewer’s signals (further referred to as CF) 42 statistical features and 16 RQA features were included. Furthermore, both PF and CF contained columns with the unique participant ID, dominant hand and the time the 10 s long sample comes from. All the features were given a unique name that are going to be used in the rest of this paper and are reported in table 2 together with a short description of each feature and mark indicating to which dataset it belongs.

### Recurrence quantification analysis

To capture and measure the repeating patterns within and between the signals, RQA was performed using the R package crqa (Moreno I., Dale, Dixon, & Nash, 2018). For extensive explanation and discussion of RQA consult (Moreno I. et al., 2018). To obtain comparable and generalizable (RQA) features a parameter optimization was performed so that the recurrence rate (RR) would fall between 3.5% and 6%. Then the median value of the parameters was used for running RQA. 2 sets of RQA parameters were obtained, one for PF samples and another for CF samples and are reported in Table 3. For PF the signal from dominant hand was compared to itself. In other words, the recurrence within the signal was identified. For CF the participant’s signal from dominant hand was compared to interviewer’s signal from dominant. Furthermore, to calculate the features relating to coordination in CF, marked with “\_diff”, the RQA features were calculated also on interviewer’s dominant hand signal compared against itself. To test that the parameter optimization was successful the mean RR was inspected. Also, logistic regression with random intercept for participant ID was performed to test whether there is a relationship between diagnosis and RR.

## Data partitioning

After the preprocessing, 25% of data were withhold from the model training in order to have independent data to validate the models. To prevent information leakage all samples from one participant were assigned to the same partition. Due to varying length of the recordings and technical problems the number of samples per participant differ. Therefore, a linear regression was used to test whether there is a significant difference in number of samples per participant between the two classes of diagnosis. Due to preprocessing and feature extraction the number of samples in PF and CF differ. The data partitions and class balance within them are reported in Table X.

## Feature Selection

To reduce the risk of overfitting, several feature selection methods were used to reduce the number of features. Although some machine-learning models are believed to have implicit feature-selection built-in, these implicit processes are often not robust enough which increases the risk of overfitting (Saeys, Abeel, & Van de Peer, 2008). We are using unsupervised feature selection instead of using domain knowledge to select the features manually because this approach was already used for feature extraction, therefore in the researchers’ opinion all the features should contain information about the diagnosis. Furthermore, the used methods reduce the risk of selection bias (Kuhn & Johnson, 2013).

First near zero variance and linear combination filters were applied to both datasets. Next, 3 different wrapper methods were used separately, so we got 3 different feature selections for each dataset. The reason behind this step is that no feature selection method guarantees the perfect performance on every dataset, it might decide to drop important features or leave features with low to no predictive power. Therefore, all feature selections were used to fit models to determine the best selection empirically *(See Evaluation)*. All feature selection was done using only the training data.

### Filtering

3 filter feature selection methods from the R package caret (Kuhn, 2018) were used on both datasets. First the zero variance and near zero variance filter was applied. As its name suggests, the zero variance filter removes features that are completely populated by one value only. The near zero variance removes features that have few unique values relative to the sample size (<10%) and/or have large ratio of the value with the highest frequency to the value with second highest frequency (>19). It is sometimes argued that excluding near zero variance might hurt the models because they might be splitting the data by the target variable perfectly (Martins, 2014). However, we argue that near-zero features in our dataset are most likely caused by the nature of the data, i.e. minimum value being mostly zero because every participant had to be still at some point during the interview. Next, a linear combination filter was applied. This method locates features that can be decomposed into linear combination of other features, i.e. mean can be decomposed into sum of data points and their number.

### Method 1 Boruta

To obtain the first selection of features a Boruta method was used. Boruta is a wrapper method which means the selection works by evaluating several models trained on different subsets of features. The R implementation of the algorithm was used (Kursa & Rudnicki, 2010). This algorithm tries to estimate the importance of the feature by creating shuffled copies of the feature (called shadow features), training a Random Forest classifier and comparing the Z-score of mean decrease accuracy of the original feature with the maximum Z-score of the shadow features. If the feature has scores higher than its shuffled copy a “hit” is recorded. We let the algorithm run 150 iterations of this process. If the feature did not score any hit after 15 iterations, it is rejected. The final output is then number if hits for all features. A potential problem of this method is the dependency of the evaluation on a model.

### Method 2 Filter correlation

The second selection of features is the result of simple filtering out correlated features. The features with absolute correlation above 0.8 were removed. The implementation of this filter method in the R package caret (Kuhn, 2018) was used.

### Method 3 Featexp

The third feature selection method used is called featexp and is implemented in Python package of the same name (Pawar, 2018). It is a model agnostic method which unlike the previous methods uses out-of-sample data to evaluate the feature importance. First, the dataset is split into train and validation data (not the validation set we set to the side at the beginning of the analysis), 50-50 ratio was used. Then population bins of numeric feature are created, by taking value intervals from the feature, i.e. first bin are all samples whose value of given feature is -1 to 0, second bin is then populated by samples with value from 0 to 1, and so on. The average of the target variable is then calculated for all population bins. This way a trend in the feature is constructed. The same process is then applied on the validation data. The feature is deemed important if the correlation of the training and validation trends lies above 0.8. This process was cross-validated with 5 folds with samples from the same participant kept within the same fold.

## Model training

5 types of models were chosen to fit to all datasets and sets of features obtained from the steps described above; model averaged neural networks (avNNet), naïve Bayes (NB), penalized logistic regression (plr), random forest (rf) and extreme boosting machine (XGBoost). That means 15 and 15 models trained on PF and CF were fitted in total. Before the data was fed to the model, all data was centered and scaled. To obtain stable performance measurement and to control for overfitting we used 30 times repeated 10-fold cross-validation with samples from the same participant kept within the same fold to prevent information leakage. This rather large number of repeats was used to account for the random folds splits with randomness. Upsampling within cross-validation was used so the no information rate was always kept at 0.5 in order to prevent any bias in the model predictions caused by different baseline probabilities. Hyperparameter tuning for all models was run for all models. Adaptive hyperparameter search based on futility analysis was used to explore the parameter space more efficiently (Kuhn, 2014). The choice of this resampling technique, however, does not allow the use of parallel processing so the actual speed-up generated by the process might not be significant. The performance metric to be maximized during the model training was area under the ROC curve. All model training was done using the R package caret (Kuhn, 2018).

## Evaluation

As the evaluation metric of all models, area under the ROC curve was chosen. However, the model evaluation had to be split into several stages because of the high number of trained models. First all models for each dataset, PF and CF, were evaluated separately, in order to select the best feature selection set. This was done by comparing the performance of the models on all resamples, that is performance calculated on the out-of-sample data during all iterations of cross-validation. With 30 times repeated 10-fold cross-validation that means we had 300 performance measurements for each model. The models were ordered by mean performance and scored from 1 to 15 in this order with 1 being the best model. These scores were then summed up to obtain the performance score of each feature selection set. Due to practicalities, we do not report the hyperparameters of all models but only their area under the curve. Only the 10 best models are reported with full model specification later. After this stage the number of models was reduced to 5 for each dataset. Then performance of these 10 models were compared in order to test the hypothesis of this paper. For this evaluation the validation dataset was used. Another test of the performance of the models was to obtain the final prediction for all participants. This was done in two ways, first the average predicted probability of all samples was taken. With the second approach weighted mean of the predicted probability was calculating, with the time variable used as weights. At the end of the evaluation we have selected the 5 best performing models. Finally, we tested the performance of the models with a mixed effect linear regression using root mean square error (rmse) as the output and the dataset as predictor and the participant as random intercept to account for the repeated measures; multiple samples from each participant. The rmse was computed for every sample using every model.

## Model interpretation

Because, most of the selected algorithms, except for the penalized logistic regression, are so-called black-box models we decided to use techniques for prediction understanding to interpret what is happening inside of the models and how the diagnosis is produced. For this purpose, the R package DALEX (Biecek, 2018) was used. Only the 5 best performing models were explored and interpreted. For the interpretation we used training and validation data separately in order to contrast the outputs of the models. This should allow us to see the goodness of fit of the models.

First to understand the model performance and what kind of errors occur in prediction the model’s distribution of residuals was calculated, compared across all 5 models and plotted as boxplots. Then we computed the feature importance scores for all models. The importance is calculated by the means of permutation. In other words, the tested feature is randomly shuffled and the change in the performance of the model is calculated. Therefore, the bigger the performance decay, the higher the importance of the feature. Because the permutation results are unstable when ran only once we ran 1000 iterations and looked at the average importance values. Next to explore the patterns driving the predictions the accumulated effects (ALE) plots of all features that showed some importance for at least one model. ALE plots are more suited for explaining our models because they deliver stable estimates even when the features are correlated. For detailed explanation of ALE plots consult Apley (2016). Finally, we used SHAP decomposition to explain 2 randomly selected participants, one control and one schizophrenic, to showcase these types of prediction-level methods for ML interpretation and its potential use in deployment of ML models to practice, e.g. psychiatrist’s office. The core element of this method are Shapley values which are a game-theory based method for determining contribution of each feature to the final prediction. For detailed discussion of the implementation of Shapley values in ML interpretability consult Lundberg and Lee (2017).

# Results

In this section the results are reported. Starting with tests confirming the preprocessing did not influence the structure of data in terms of predictability of diagnosis. Next, the best set of selected features is selected for both PF and CF. Then the evidence for the main hypothesis is reported including performance evaluation of all models trained on the best feature selection sets. Finally, the interpretations of the best performing models, chosen based on the hypothesis testing, are reported.

## Preprocessing tests

Linear regression showed that there is no systematic difference in number of samples per diagnosis in neither of the data partitions. The results are reported in Table X. The average values of RR were acceptable with 8.37 and 7.05 in PF and CF respectively. Neither of the logistic regressions indicated that RR is significant predictor of diagnosis (Table X). Therefore, the RQA features could be used for the model training.

## Feature selection

The results of zero variance and near zero variance filters are reported in figure X for both PF and CF. 2 and 3 features were removed from PF and CF respectively.The features removed due to linear combination filter are reported in table X together with their linear combination of origin. After applying these 3 filters there are 23 and 48 features left in PF and CF respectively.

### Participant feature selections

Using the Boruta method 23 features were selected (Table X). The importance estimated are plotted in figure X. Using the correlation-filter method 18 features were selected (Table X). The featexp method selected 10 features (Table X).

### Coordination feature selections

Using the Boruta method 35 features were selected (Table X). The importance estimated are plotted in figure X. Using the correlation-filter method 32 features were selected (Table X). The featexp method selected 10 features (Table X).

## Evaluation of the feature selections

Auc and score of all models trained on all feature selections of PF is reported in table X. The same way the auc and score of all CF models is reported in table X. For better understanding the figure X shows the same information with the standard deviation of auc added for both PF and CF models. The final scoring of each feature selection is reported in table X. The Boruta selection was chosen as the best PF selection. The featexp selection was chosen as the best CF selection.

## Participant vs. coordination models on validation data

The full architecture of the 10 best models is reported in table X. The performance of all 10 models on the validation data is reported in table X. In figure X the performance of the models is plotted for easier comparison. The performance of the models calculated on the averaged predictions over participants is reported in table X. The performance of the models calculated on the weighted average predictions is reported in table X. The results of the mixed effect model testing the performance the models in terms of rmse is reported in table X.

## Model interpretation

The distributions of residuals of all 5 models on training and validation data are plotted in figure X and X. The feature importance based on validation data is plotted in figure X. Due to practicalities ALE plots of only those features discussed in the discussion are reported here. The ALE plots of the rest of the features are included in the appendix. An example of using SHAP decomposition to explain an individual prediction is reported in figure X.

# Discussion

## Discussion of the results

All used tests of predictive power of the models support the hypothesis; the models trained on features relating to coordination of the interlocutors perform better on the validation data better than models using features computed only from the participant’s signal. They perform better not only in terms of the main evaluation metric that the models were trained to maximize; the area under the curve; but also, in terms of sensitivity and specificity. The support could be considered even stronger as the coordination models outperform the participant models with significantly lower number of features, 10 and 23 features respectively. Important to note is that the results of the averaged predictions do not differ much when simply averaged or weighted by the time of the sample. From the view of the interpersonal synergy theory the coordination system should emerge gradually over time and therefore the prediction power should be higher in more established system, however we do not see this pattern in the data.

Moving on to the interpretation of the coordination models we can see that having hold out the validation data is quite crucial to independent evaluation of the models. As could be expected, the residuals of the models are considerably larger when calculated on the validation data. Especially striking difference is visible in case of the random forest. This model quite likely overfit to the data, learning too concrete patterns from the training data which are not generalizable to previously unseen data. The feature importance plot shows another strange behavior of random forest and also of naïve Bayes when some of the intervals go into the opposite direction. In other words, the performance of these models improves when these features are randomly shuffled. At this moment, we have no explanation for random forest of what the cause might be. In case of naïve Bayes, it might be because of the violation of the feature independence assumption. The most important features seem to be the difference of sums (DS) of all movements, difference of interquartile range (DIR) and therefore it is worthwhile to look at their ALE plots calculated from validation data.

The pattern in the DS learned by all models is the exact opposite of what the research described in the introduction suggested. When the participant is moving more than the interviewer, i.e. DS > 0, the probability of being schizophrenic is increases. Similarly, the pattern in the DIR contradicts the previous research. It suggests that when we ignore the very small and very large movements and look only at the “average” movements the schizophrenics tend to gesture more than the psychologist. Important thing to note in this ALE plot is the first points (from the left) on each lines that are very far from the other points. This suggests that value of DIR this low is very rare and the models did not see such value before which leads to strange behavior of the predictions, e.g. the random forest would increase the schizophrenia probability while the rest of the models would decrease it. In other words, we should ignore that part of the ALE line when interpreting the models. It might indicate a space for model improvement but that is beyond the scope of this thesis.

We offer two post hoc explanations of these patterns, that should be treated very skeptically as they are only speculations. First is that the patient is the one who is compensating the lack of nonverbal communication. In the recordings of the interviews the psychologist can be heard writing something down which could lead to the patient feeling the need to compensate. Second explanation might be that the psychologist, aware of the diagnosis, is giving the patient more space to express and neglecting her appropriateness of communication. In order to resolve this contradiction between our results and previous research, designing a new experiment would be appropriate.

In the introduction, we mentioned that for a model to be deployed in real-life scenario, e.g. in psychiatric office, its predictions have to be justifiable in an easy manner. Figure X presents a prototype of such justification. This way the psychologist would have access to more information which could improve the decision making. Using hand gesture for such predictive system is practical as the setup for data collection is relatively simple and the clinical interview has to be performed in all cases as the current diagnostical system demands.

## Interviewer bias

One of the biggest limitations of this study is that the psychologist is aware of the participant’s condition from the very beginning of the interview. Therefore, it could be argued that the psychologist’s behavior is biased. However, should this system be applied in real-life case, this bias would not be present as the psychologist should start the interview thinking the patient is healthy. Currently it is not possible to estimate the risk this bias poses. The best way to test whether the predictive performance is consequence of this bias, would be running the same data collection setup as used in this with the difference of using randomly recruited people as interviewers instead of a trained professional.

## Retraining the models

Now that we know that for the models to perform well, RQA based features are not required. In this case the use of RQA has become a limitation of the study. The RQA cannot be applied to all actigraph signals we had. The data loss happens because the optimized parameters do not allow the computation of recurrence on some samples. More data was lost due to high recurrence rate in some samples reaching up to 99% which meant removing these samples from further analysis. The RQA cost us approximately 60% of the original sample size. Therefore, it is reasonable to expect a change in the performance of the models if they were retrained using the full sample size.

# Conclusion

The results support the hypothesis that social function can be better explained as interpersonal system emerging form the interaction rather than attributing the social function to the individual. Furthermore, the predictive models reached reasonable performance with small number of features and limited sample size. The patterns learned from the data are in contradiction with previous research on the topic of interlocutor coordination of hand gestures. Further research is required to resolve the conflict. Two limitations of the study were identified, and plausible solutions suggested.

# References

American, Psychiatric, & Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*: American Psychiatric Pub.

Andrey, A. (2018). Soundgen: An open-source tool for synthesizing nonverbal vocalizations. *Behavior Research Methods, 1-15*. doi:10.3758/s13428-018-1095-7

Apley, D. W. (2016). Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *ArXiv e-prints*.

Biecek, P. (2018). DALEX: explainers for complex predictive models. *ArXiv e-prints*.

Bora, E., & Pantelis, C. (2016). Social cognition in schizophrenia in comparison to bipolar disorder: a meta-analysis. *Schizophrenia research, 175*(1-3), 72-78.

Borchers, H. W. (2018). pracma: Practical Numerical Math Functions (Version R package 2.1.8). Retrieved from <https://CRAN.R-project.org/package=pracma>

Brunet-Gouet, E., & Decety, J. (2006). Social brain dysfunctions in schizophrenia: a review of neuroimaging studies. *Psychiatry Research: Neuroimaging, 148*(2-3), 75-92.

Deacon, T., & Poeppel, D. (1997). The symbolic species: The co-evolution of language and the brain. *Nature, 388*(6644), 734-734.

Delude, C. M. (2015). Deep phenotyping: The details of disease. *Nature, 527*, S14. doi:10.1038/527S14a

Doshi-Velez, F., & Kim, B. J. a. p. a. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint*.

Fried, E. I., & Nesse, R. M. (2015). Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR\* D study. *Journal of affective disorders, 172*, 96-102.

Fusaroli, R., Konvalinka, I., & Wallot, S. (2014). Analyzing social interactions: the promises and challenges of using cross recurrence quantification analysis. In *Translational recurrences* (pp. 137-155): Springer.

Fusaroli, R., Rączaszek-Leonardi, J., & Tylén, K. (2014). Dialog as interpersonal synergy. *New Ideas in Psychology, 32*, 147-157. doi:<https://doi.org/10.1016/j.newideapsych.2013.03.005>

Fusaroli, R., Weed, E., Simonsen, A., & Bliksted, V. F. (2013). *Non-linear dynamics of speech and voice in schizophrenia.* Paper presented at the Neurobiology of Language.

Gallese, V., & Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive neuropsychology, 22*(3-4), 455-479.

Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA)*.

Howes, C., Lavelle, M., Healey, P. G., Hough, J., & McCabe, R. (2017). *Disfluencies in dialogues with patients with schizophrenia.* Paper presented at the CogSci.

International statistical classification of diseases and related health problems. (2016). Geneva, Switzerland: World Health Organization.

Kuhn, M. (2014). Futility analysis in the cross-validation of machine learning models.

Kuhn, M. (2018). caret: Classification and Regression Training (Version R package version 6.0-80). Retrieved from <https://CRAN.R-project.org/package=caret>

Kuhn, M., & Johnson, K. (2013). An Introduction to Feature Selection. In *Applied Predictive Modeling* (pp. 487-519). New York, NY: Springer New York.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software, 36*(11), 1-13.

Lavelle, M., Healey, P., & McCabe, R. (2012). Is Nonverbal Communication Disrupted in Interactions Involving Patients With Schizophrenia? (Vol. 39).

Lavelle, M., Healey, P. G. T., & McCabe, R. (2014). Participation during First Social Encounters in Schizophrenia. *PLOS ONE, 9*(1), e77506. doi:10.1371/journal.pone.0077506

Lavelle, M., Howes, C., Healey, P., & McCabe, R. (2013). *Speech and hand movement coordination in schizophrenia*.

Lundberg, S. M., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions.* Paper presented at the Advances in Neural Information Processing Systems.

Martins, T. G. (2014). Near-zero variance predictors. Should we remove them? Retrieved from <https://tgmstat.wordpress.com/2014/03/06/near-zero-variance-predictors/>

Moreno I., C., Dale, R., Dixon, J. D., & Nash, J. (2018). crqa: Cross-Recurrence Quantification Analysis for Categorical and

Continuous Time-Series (Version R package version 1.0.9). Retrieved from <https://CRAN.R-project.org/package=crqa>

Nielsen, D. (2016). Tree Boosting With XGBoost-Why Does XGBoost Win" Every" Machine Learning Competition? (Master’s Thesis), NTNU,

Parola, A., Simonsen, A., Bliksted, V., & Fusaroli, R. (2018). T138. Acoustic Patterns In Schizophrenia: A Systematic Review And Meta-analysis. *Schizophrenia Bulletin, 44*(1), S169-S169.

Pawar, A. (2018). My secret sauce to be in top 2% of a kaggle competition: Towards Data Science. Retrieved from <https://towardsdatascience.com/my-secret-sauce-to-be-in-top-2-of-a-kaggle-competition-57cff0677d3c>

Puiatti, A., Mudda, S., Giordano, S., & Mayora, O. (2011). *Smartphone-centred wearable sensors network for monitoring patients with bipolar disorder.* Paper presented at the Engineering in Medicine and Biology Society, EMBC, 2011 annual international conference of the IEEE.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *Why should i trust you?: Explaining the predictions of any classifier.* Paper presented at the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.

Ripley, B. D. (2007). *Pattern recognition and neural networks*: Cambridge university press.

Saeys, Y., Abeel, T., & Van de Peer, Y. (2008). *Robust Feature Selection Using Ensemble Feature Selection Techniques*, Berlin, Heidelberg.

*Schizophrenia Fact sheet N°397*. (2015). World Health Organization.

Schneider, K. (1959). Clinical psychopathology (Trans. by MW Hamilton).

Simonsen, A., Fusaroli, R., Skewes, J. C., Roepstorff, A., Campbell-Meiklejohn, D., Mors, O., & Bliksted, V. (2018). Enhanced Automatic Action Imitation and Intact Imitation-Inhibition in Schizophrenia. *Schizophrenia Bulletin*, sby006-sby006. doi:10.1093/schbul/sby006

Tsuang, M. T., Nossova, N., Yager, T., Tsuang, M. M., Guo, S. C., Shyu, K. G., . . . Liew, C. (2005). Assessing the validity of blood‐based gene expression profiles for the classification of schizophrenia and bipolar disorder: A preliminary report. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 133*(1), 1-5.

Tzur Bitan, D., Grossman Giron, A., Alon, G., Mendlovic, S., Bloch, Y., & Segev, A. (2018). Attitudes of mental health clinicians toward perceived inaccuracy of a schizophrenia diagnosis in routine clinical practice. *BCM Psychiatry, 18*(1). doi:10.1186/s12888-018-1897-2

Woodberry, K. A., Giuliano, A. J., & Seidman, L. J. J. A. J. o. P. (2008). Premorbid IQ in schizophrenia: a meta-analytic review. *The American Journal of Psychiatry, 165*(5), 579-587.

# Appendix