

Emil Trenckner Jessen (EJ)  
(au604547@uni.au.dk)  
Cognitive Science, Aarhus University  
Jens Chr. Skous Vej 2,  
8000 Aarhus, Denmark

Voice atypicalities in Schizophrenia; replicability of machine learning approaches

Contents

[1. Introduction 2](#_Toc57831910)

[1.1 Schizophrenia and biomarkers 2](#_Toc57831911)

[1.1.1 Schizophrenia and voice atypicalities 2](#_Toc57831912)

[1.2 Machine learning for detection of acoustic patterns 2](#_Toc57831913)

[1.2.1 Prospects of machine learning in classifying schizophrenia 2](#_Toc57831914)

[1.2.2 Current limitations in the literature 3](#_Toc57831915)

[1.3 Alleviating current limitations 3](#_Toc57831916)

[1.3.1 Through replications and conservative ML implementation 3](#_Toc57831917)

[1.3.2 A rigorous ML pipeline 3](#_Toc57831918)

[1.3.3 Purpose of paper 4](#_Toc57831919)

[2. Methods 4](#_Toc57831920)

[2.1 Pipeline 4](#_Toc57831921)

[2.2 Literature search and choice of replication 6](#_Toc57831922)

[2.3 Data 7](#_Toc57831923)

[2.3.1 Data sources 7](#_Toc57831924)

[2.3.2 Participants 7](#_Toc57831925)

[2.3.2 Procedure/task 8](#_Toc57831926)

[2.4 Preprocessing 8](#_Toc57831927)

[2.4.1 Cleaning of audio files 8](#_Toc57831928)

[2.4.2 Feature extraction from audio files 8](#_Toc57831929)

[2.5 Partitioning 9](#_Toc57831930)

[2.6 Normalization 9](#_Toc57831931)

[2.5 Feature selection using LASSO 10](#_Toc57831932)

[2.5.1 Motivation for using LASSO 10](#_Toc57831933)

[2.5.2 What is L2 regularization? 10](#_Toc57831934)

[2.5.4 Feature selection 11](#_Toc57831935)

[2.6 Parameter tuning, model training and model testing 12](#_Toc57831936)

[2.7 Evaluation metrics 12](#_Toc57831937)

[3. Results 13](#_Toc57831938)

[4. Discussion 13](#_Toc57831939)

[4.1 Results and comparison to original study 13](#_Toc57831940)

[4.1.1 Performance (comparison) 13](#_Toc57831941)

[4.1.2 Methods (comparison) 14](#_Toc57831942)

[4.2 Pipeline 15](#_Toc57831943)

[4.2.1 How did an implementation of pipeline in this replication work out? 15](#_Toc57831944)

[4.2.2 Insights on problems in research derived from doing a conservative replication 16](#_Toc57831945)

[4.3 Further research 17](#_Toc57831946)

[4.3.1 Taking everything into account: How does the pipeline help further research? (Wrap-up) 17](#_Toc57831947)

[5. Conclusion 17](#_Toc57831948)

[6. References 18](#_Toc57831949)

[8. Appendix 19](#_Toc57831950)

**Abstract**

s

**Keywords:** Schizophrenia, Voice, Machine Learning, SVM

# 1. Introduction

## 1.1 Schizophrenia and biomarkers

### 1.1.1 Schizophrenia and voice atypicalities

1. Schizophrenia impairments/symptoms
   1. Alogia
   2. Blunting of affect
   3. Emotional impairment
      1. Unable to detect emotion from voice
2. Important to diagnose early
3. (Brief) history on voice atypicalities
4. Voice as a biomarker for diagnosis
5. Previous findings on voice
   1. Qualitative perceptual ratings
      1. Robust differences betwn TD and SZ
      2. Expensive
   2. Quantitative perceptual ratings
      1. Fewer robust differences betwn TD and SZ
      2. Varying effect sizes and sometimes direction
   3. (Brief) Machine learning
      1. Promising results
      2. Objective
      3. Well-suited for clinical application

## 1.2 Machine learning for detection of acoustic patterns

### 1.2.1 Prospects of machine learning in classifying schizophrenia

1. Meta (link to previous section)
   1. What is machine learning in classifying schizophrenia from voice?
      1. Will not go into other ML things (beyond scope of paper)
         1. Cite paper with gesticulation + voice or others
2. What does it allow for?
   1. Finding relevant acoustic features (feature selection)
   2. Complex analysis of multiple features in conjunction
   3. Cheap classification (as opposed to qualitative)
3. Promising findings (high accuracy in many studies)
4. A lot of possibilities of practical application
5. Less interpretability

### 1.2.2 Current limitations in the literature

1. Differences in methods, method quality and levels of transparency and documentation
2. Large difference in performance
   1. Perhaps overfitting
3. Lack of replications
   1. Promising results
   2. No validation between datasets
   3. No information on robustness
   4. Language and diagnosis differences

## 1.3 Alleviating current limitations

### 1.3.1 Through replications and conservative ML implementation

1. How replications alleviate limitations
2. How conservative ML implementation alleviates limitations
   1. Required for good original studies
   2. Required for having the replications be useful
   3. Required for having the replications be possible at all
3. But what is a conservative ML implementation?

### 1.3.2 A rigorous ML pipeline

1. What is a pipeline?
2. Pipeline proposal:
   1. Data acquisition
   2. Preprocessing
   3. Data partitioning
   4. Feature scaling
   5. Feature selection
   6. Model tuning (training, tuning and testing cycle)
   7. Validation (and evaluation)
   8. (Reflection + proper documentation)

### 1.3.3 Purpose of paper

1. Thesis statement
   1. In other words:
      1. Provide pipeline
      2. Show example of implementation

# 2. Methods

## 2.1 Pipeline

1. This project attempts to follow and exemplify good ML practice mentioned in in introduction
2. The rest of methods section will go into detail, but here an overview/summary will be provided
3. Pipeline
   1. Data acquisition
   2. Data preparation/preprocessing
   3. Partitioning
   4. Feature scaling and reduction
   5. Model training, testing and parameter tuning
   6. Validation with proper use of information on evaluation

The replication of this paper follows and provides an exemplification of the use of a rigorous pipeline, following the overall principles presented in the introduction (see section 1.3.2). The rest of the methods section will provide a detailed description of the course of action taken to replicate the paper by \* Chakraborty et al. \*. However, to provide an overview of the process and showcase how it followed our proposed pipeline, a short summary will be provided along with two figures. One which attempts to visualize the pipeline (figure 1) and one which attempts to visualize the complex multi-leveled process of partitioning of the data (figure 2).

Diagram

Description automatically generated

Figure 1.

*An overview of the proposed pipeline. Purple boxes refer to the general pipeline whereas the yellow refer to the specifics used in this replication.*

The pipeline for this replication is divided up into 8 steps which are as follows:

**1) Data acquisition.** Recorded voice data from 4 studies was acquired. **2)** **Data preprocessing.** Prior to this study, all data had gone through a cleaning process and all features extracted in the original experiment were extracted from the data. **3) Partitioning.** The data was then partitioned into a training a holdout set of 80% and 20%, respectively. **4) Feature scaling and selection.** Features were min./max. normalized. **5) Feature selection.** The training data was split up into 5 folds. These 5 folds were used to create 5 splits – with each split having a training (4 out of 5 folds) and a testing set (remaining 1 out of 5 folds). All training sets had their features L2 regularized; resulting in 5 feature sets of only the most relevant features for classification of schizophrenia. This process did in other words produce 5 splits of training and testing data. It also produced 5 feature sets. Each of the feature sets had been selected on the basis of 4/5th of the full training data, which meant that the remaining 1/5th could be used for testing (for an overview, see figure 2).

A picture containing timeline

Description automatically generated

Figure 2.

*A visualization of the data structures used for this ML replication. Shows the process of partitioning and dividing the data up into multiple folds and outlines which sets were used for what.*

**6), 7), 8) Model training, parameter tuning, model testing.**

5 SVM linear kernel models were then constructed to classify patients from controls. Each of these models were fit on training sets (4/5th of the full training data) and tested on the matching test set (last 1/5th). These models used the parameters from the training set had had been L2 regularized. The predictions were then evaluated based on their classification performance and C and Gamma parameters were tuned. After tuning, the models were then tested again – repeating this process until needs of satisfactory performance were met.

**9) Validation on holdout set.** Finally, the 5 models were tested on the holdout set. An ensemble model was also constructed and test. This model predicted the holdout data, by using the majority vote as a prediction. Performance on the holdout set was then evaluated for the 6 models with the relevant metrics. Performance of the sexes separately was also calculated to allow for insights into potential ML biases.

## 2.2 Literature search and choice of replication

A literature search for papers, dissertations and unpublished manuscripts was conducted for finding a paper to replicate. The complete list of papers listed in the meta-analysis by Alberto et al. in 2019 (Alberto et al., 2019) was manually screened – first by title and since by content. As their search was last updated as of April 12 2018, the search was continued from that date and forward in time by the use of search using Google Scholar on the Sep 15 2020, using the same search terms (schizo\* AND machine learning AND prosody OR inflection OR intensity OR pitch OR fundamental frequency OR speech rate OR voice quality OR acoustic OR intonation OR vocal).

The manual search explored the papers by the author, looking for papers that 1) were transparent and well-documented, 2) were thorough in applying proper machine learning methods, 3) had larger amounts of data. The study by Chakraborty et al. from 2018 was chosen for replication after assessing the literature on these parameters (Chakraborty et al., 2018).

## 2.3 Data

### 2.3.1 Data sources

The data used in this paper consists of speech recordings gathered from 3 published studies (Beck et al., 2020; Bliksted et al., 2014, 2019) and an unpublished study by Vibeke Bliksted.   
Although the data was acquired in separate studies the speech data has several qualities which makes it suitable for combining into a single study:

Participants from all studies went through the same tasks; namely the Frith Happé animations task (Abell et al., 2000). All participant went through 8 such trials, except for in the study from 2015 by Bliksted et al., where the they went through 10 trials (Bliksted et al., 2014).

Moreover, recording equipment and recording setting was constant within study, but unique across studies. This results in data corpora of diverse speech recordings suitable for testing whether implementation of a certain machine learning algorithm proves to be versatile in its predictions across data sets.

### 2.3.2 Participants

222 Danish participants were included in this study. Out of the 222 participants 106 were clinically diagnosed with schizophrenia by trained psychiatrists in accordance with the standards of ICD-10 DCR (Zivetz, 1992). Patients were recruited through OPUS, Clinic for people with schizophrenia, Aarhus University Hospital Risskov.  
The patient group was originally matched one-to-one with healthy control subjects (N = 116), using the following criteria: age, sex, handedness, ethnicity, community of residence and parental social economic status (based on the highest parental education and expected parental income according to Statistics Denmark regarding wages) and educational level (based on the last commenced education) (*Statistics Denmark*, n.d.). Healthy control subjects were recruited via advertisements in four local newspapers. All participants in this group (and their first-degree relatives) had no history of any psychological disorders. Although the control group was originally matched one-to-one with the patient group, 14 patients and 4 controls were excluded due to poor recording quality or other similar factors. This explains the uneven number of participants within each group. For further information on participants, see table 1.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Study | N() | Diagnosis | N(Females) | N(Males) | Mean(Age) | SD(Age) | Range(Age) |
| Beck et al., 2020 | 70 | SZ | 16 | 18 | 22.8 | 3.13 | 18-31 |
| TD | 17 | 19 | 22.7 | 3.19 | 18-30 |
| Bliksted et al., 2014 | 46 | SZ | 6 | 17 | 23.3 | 3.94 | 18-33 |
| TD | 7 | 16 | 23.7 | 3.61 | 18-34 |
| Bliksted et al., 2019 | 48 | SZ | 11 | 8 | 40.8 | 12.4 | 20-61 |
| TD | 13 | 16 | 37.5 | 13.1 | 21-62 |
| Bliksted et al., n.d. | 58 | SZ | 12 | 18 | 24.8 | 3.66 | 18-31 |
| TD | 13 | 15 | 24.4 | 4.65 | 18-34 |
| Total | 106 | SZ | 45 | 61 | 26.7 | 9.02 | 18-61 |
| 116 | TD | 50 | 66 | 26.7 | 9.22 | 18-62 |

Table 1:

*Demographic data on the sex and diagnosis within each of the different studies. N means number and SD standard deviation.*

### 2.3.2 Procedure/task

The participants went through the Frith Happé animations task. This task consisted of watching a 2D top-view video of animated triangles. There were two distinct triangles; one large red and one small blue, both of which moved around on the screen and most videos furthermore contained an enclosure in the center of the video. There were three conditions with multiple videos for each condition:

**1. Random movement sequences.** There was no obvious interaction between the triangles and movement appears random. **2. Goal-directed (G-D) movement sequences.** An interaction between the triangles in which actions are directed toward each other in order to achieve specific goals.

**3. Mental interaction (ToM)**. An interaction between the triangles involving the manipulation of the emotions and thoughts of one triangle by the other. After watching an animation from one of these conditions, the participants were interviewed and asked to describe what happened in the animation. Each description of a trial thus ended up as a single .wav file.

## 2.4 Preprocessing

### 2.4.1 Cleaning of audio files

The cleaning of the audio files was carried out by Ludvig Olsen in 2018 (Olsen, 2018)  
The audio files were then converted to 16-bit .wav files, with a sample rate of 16k. They were subsequently denoised by stacking multiple instances of the Voice De-noise and De-hum tools in the iZotope RX 6 audio editor (iZotope Inc., 2018). A small equalizer tilt was applied at 1085Hz with the Fabfilter Pro-Q2 equalizer to bring more brightness to the signal (FabFilter Software Instruments, 2018). The signal was normalized to peak at -1dB both before and after the cleaning steps.

### 2.4.2 Feature extraction from audio files

The toolkit openSMILE 2.3.0 was used for extracting the features needed for the SVM classification algorithm. From within the openSMILE software package, the base-set configuration file of emotion recognition features called ‘emobase’ was chosen for feature extraction.

The feature set specified by emobase contains 988 features used for emotion recognition:

Intensity, Loudness, 12 MFCC’s, F0 Pitch, Probability of voicing, F0 envelope, 8 LSFs (Line

Spectral Frequencies), Zero-Crossing Rate. Delta regression coefficients are then computed from all these previously mentioned low-level descriptors (LLD). Both the LLDs and their delta coefficients are smoothed by a moving average window that filters with a window size of 3 seconds. Furthermore, the following functionals are applied to the LLDs and the delta coefficients:

Max./Min. values and their respective relative position within input, range, arithmetic mean, 2 linear

regression coefficients and linear and quadratic error, standard deviation, skewness, kurtosis,

quartile 1-3, and 3 inter-quartile ranges.

This results in the feature set consisting of 988 features. In other words; 26 LLDs, a delta regression coefficient for each LLD and 19 functionals for each of the LLDs and for each of the delta regression coefficients (26 \* 2 \* 19 = 988). The process of feature extraction was executed on each of the speech recordings, yielding a single feature vector for each trial of each participant.

## 2.5 Partitioning

To be able to evaluate the performance of the model the dataset was partitioned into a training set and a holdout set consisting of 80% and 20% of the total data, respectively. The partitioning was carried out using the package groupdata2 and was done semi-randomly (Olsen, 2020). The partitioning kept each participant ID only within one of the two resulting training set and holdout set. This prevented leakage of information from the training set to the holdout set, which otherwise would have led to overfitting and as a result an unprecise evaluation of out-of-sample performance. Moreover, to avoid a skewed distribution of sex or diagnosis between the training and the holdout set (e.g. ending up with only males/controls in the holdout set as a result of a random partitioning), sex and controls/patients were evenly distributed in the partitioning. The constituents of the holdout set were sound files for each trial from 11 female controls, 10 female patients, 12 male controls and 11 male patients. A properly balanced holdout set ensures that performance across sexes and diagnosis can be accessed without too much statistical uncertainty (e.g. when calculating a model’s accuracy for females, with only predictions on a single female).

## 2.6 Normalization

All feature parameters were normalized using the min-max feature scaling formula in order to achieve a dataset with a common scale without losing information or distorting differences in the range of values.



To avoid overfitting as a result of carrying data from the test set to the training set, the normalization was carried out separately for the training and the testing set. The scaling used the min. and the max. value for each feature, only from the training set, both for the training and for the testing set. This had the advantage of having both the training and the test features on the same scale, while not letting information from the test set flow to the training set and is common practice when applying most machine learning algorithms.

## 2.5 Feature selection using LASSO

### 2.5.1 Motivation for using LASSO

As the 988 acoustic features from the ‘emobase’ package were originally designed to distinguish emotions from speech, some of the features were bound to be redundant for the purpose of distinguishing between patients and controls. As a measure to counterfeit this, a rigorous feature selection method was applied to rid the model of superfluous features. This was done in order to simplify the model and thereby reduces both complexity, computational power needed to run the model and in order to improve both predictive power and interpretability of the classifier.

Feature selection was done using L2 regularization, also called the Least Absolute Shrinkage and Selection Operator (LASSO) analysis regression. To carry out this process, the ‘glmnet’ R Package was utilized for the purpose of this paper. (Friedman et al., 2010)

Although the parameters could have been regularized using Ridge or ElasticNet, LASSO regularization has the advantage of being able to shrink irrelevant parameters all the way to zero – as opposed to Ridge regularization. Elastic net is a combination of Ridge and Lasso and would therefore be a compromise between the two. The shrinking of parameter estimates to zero is beneficial given the many features that are unrelated to the distinction between schizophrenia and healthy individuals.

### 2.5.2 What is L2 regularization?

This method optimizes beta estimates for all parameters not only through misclassification error but also adding a L2 regularization term. The latter adds a penalty to each beta estimate on the basis of a lambda value multiplied with the beta estimate.

In other words; performing L2 regularization means fitting a LASSO regression model and thus finding the optimal beta values for all parameters using the loss function seen below.

The loss function used for finding parameter estimates using LASSO:  


Since this method requires a lambda value (λ), the optimal lambda value also had to be found. The lambda value producing the minimum value in the loss function (lambda.min) was first computed. This was done by testing a range of lambda values using 5-fold cross-validation. Subsequently the lambda value resulting in the fewest number of parameters within 1 SE from the lambda.min was chosen (lambda.1se). Although lambda.min has the lowest level of misclassification, lambda.1se has the advantage of acknowledging the fact that the fits are estimated with some error (Friedman et al., 2010). This process thus generates a list of parameter estimates. Those that have not been shrunken to zero are selected as relevant features for predicting patients from controls. For a visualization, see figure 3.



Figure 3:

*A range of lambda values (x-axis) and the resulting 1) misclassification error, and 2) number of features (seen at the top). From left to right, the dotted lines represent lambda.min and lambda.1se, respectively.*

### 2.5.4 Feature selection

The training data was partitioned into 5 folds, and thus also 5 splits (see fig. 2). The previously mentioned L2 regularization was carried out on 4/5th’s of each of these splits, resulting in 5 different feature sets (see appendix x\* for list of these feature sets). An illustration of the feature selection for a single split, can be seen below in figure 4.

Graphical user interface

Description automatically generated

Figure 4:

*Figure showing the process of feature selection for feature set 1:*

*The training data is divided up into 5 folds. One fold is then excluded (yellow). Using cross-validation, the LASSO regression fit for a specific lambda value is then computed with each of the folds being omitted once. The misclassification error for each of these fits is then accumulated and stored. The process is then reiterated using a new lambda value from the lambda grid, until all errors from all relevant lambda values have been obtained. This entire procedure is then repeated for each of the 5 splits.*

## 2.6 Model training, testing and parameter tuning

Using the 5 training sets and the appertaining feature sets for each split, 5 SVM linear kernel classifier models were constructed. The models were fit on the trainings sets, only using the appertaining feature sets. SVM classifiers were then tested on the appropriate test sets (the model fit on training set n was tested on test set n) and performance was evaluated based on relevant metrics (see section 2.7). After testing, the models were tested again using a self-specified range of C-parameters around default (1) to see if they allowed for better predictions. The default C-parameter of 1 was found optimal for classification.

The 5 models were then implemented into an ensemble model. This model merely predicted using the majority vote of the 5 sub-models. If for example 3 out of 5 models predicted ‘schizophrenia’ then this was also the vote of the ensemble model.

## 2.7 Evaluation metrics

For evaluating the performance of the models, several metrics conveying information about the classification will be provided. Precision (positive predictive value) is the ratio between true positives and all positive predictions. Recall on the other hand is the ratio of positives that were correctly classified. Although both precision and recall are typically only provided for the model as a whole, additional information can be acquired be calculating them for each class (i.e. getting ratios for both the patient and the control group). An f1-score account for the fact that precision and recall oftentimes will be inversely correlated. A such score gives the harmonic mean of precision and recall and gives an overall understanding of the classification performance for each class. By calculating the arithmetic mean of the two F1-score for a model, a single score provides clear insight into classification performance.   
Accuracy – the percentage of correct classifications - gives an intuitive impression of the performance and is regarded common practice. However, it can often be misleading (e.g. when evaluating performance on unbalanced data). By providing information about the baseline accuracy it is possible to compare accuracy, since baseline accuracy depicts the accuracy rate of a model that merely predicted the majority group.   
Moreover, confusion matrices will be provided as they convey the whole picture of evaluation and provide all the information needed for any other evaluation metrics to be calculated.



*Where,   
tp, fp, tn, fn, refers to true positives, false positives, true negatives, false negatives, while i and N refers to class and number of classes respectively.*

# 3. Results

This section presents the performance of the machine learning models when predicting various parts of the full data. A crude overview of the performance of the 5 models on the various test sets is given in table 2. An in-depth look at the ensemble models performance; both for controls and for the patient group is provided in table 3. The latter also provides insight into performance differences between the sexes. Finally, confusion matrices (table 4, 5 and 6) provide the necessary details that would underlie calculations for any and all additional performance metrics.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Testing set** | **Training and feature set** | **Macro avg. F1-score** | **Accuracy** | **Baseline accuracy** |
| Train 1 | Train 1 | 0.896 | 89.64% | 53.05 |
| Train 2 | Train 2 | 0.930 | 93.03% | 51.52 |
| Train 3 | Train 3 | 0.897 | 89.73% | 52.21 |
| Train 4 | Train 4 | 0.899 | 89.91% | 51.89 |
| Train 5 | Train 5 | 0.898 | 89.85% | 51.80 |
|  | | | | |
| Test 1 | Train 1 | 0.687 | 68.68% | 51.85 |
| Test 2 | Train 2 | 0.630 | 63.05% | 54.34 |
| Test 3 | Train 3 | 0.678 | 67.84% | 51.62 |
| Test 4 | Train 4 | 0.613 | 61.31% | 52.94 |
| Test 5 | Train 5 | 0.658 | 65.80% | 53.29 |
|  | | | | |
| Holdout | Train 1 | 0.644 | 64.44% | 51.87% |
| Train 2 | 0.652 | 65.19% | 51.87% |
| Train 3 | 0.735 | 73.51% | 51.87% |
| Train 4 | 0.740 | 74.05% | 51.87% |
| Train 5 | 0.716 | 71.64% | 51.87% |
| **Ensemble (majority vote of set 1:5)** | **0.703** | **70.32%** | **51.87%** |

Table 2:

*Prediction performance for all 5 SVM linear kernel models, on various testing data.*

*Within-sample prediction performance can be seen in the first 5 rows, while 5-10 depicts performance tested on the 5 test sets. Finally, the performance for the models’ predictions on the holdout set and the majority decision vote can be seen in the bottommost 6 rows.*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Test set | Model | Sex | Acc. | Baseline acc. | Class | Precision | Recall | F1-score |
| Holdout | Ensemble | Male | 70.62% | 52.58% | SZ | 0.664 | 0.772 | 0.714 |
| TD | 0.759 | 0.647 | 0.698 |
| Female | 70.00% | 51.11% | SZ | 0.689 | 0.705 | 0.697 |
| TD | 0.711 | 0.696 | 0.703 |
| Both | 70.32% | 51.87% | SZ | 0.675 | 0.739 | 0.706 |
| TD | 0.734 | 0.670 | 0.700 |

Table 3:

*Performance of the ensemble model - within both the sexes and diagnosis. TD stands for typically developed (the control group) while SZ represents the schizophrenia group.*

|  |  |  |  |
| --- | --- | --- | --- |
| N = 374 | Predicted group | | |
| True group |  | TD | SZ |
| TD | 130 | 64 |
| SZ | 47 | 133 |

Table 4:  
*Confusion matrix for the ensemble model predictions*

|  |  |  |  |
| --- | --- | --- | --- |
| N = 180 | Predicted group | | |
| True group |  | TD | SZ |
| TD | 64 | 28 |
| SZ | 26 | 62 |

Table 5:

*Confusion matrix for the female ensemble model predictions*

|  |  |  |  |
| --- | --- | --- | --- |
| N = 194 | Predicted group | | |
| True group |  | TD | SZ |
| TD | 66 | 36 |
| SZ | 21 | 71 |

Table 6:

*Confusion matrix for the male ensemble model predictions*

# 4. Discussion

## 4.1 Results and comparison to original study

### 4.1.1 Performance (comparison)

1. Performance of models on test
   1. F1-score (and short mention of accuracy)
      1. F1-score for model overall
      2. F1-scores for patients and controls respectively
   2. Precision + recall
   3. Between sexes
      1. Well balanced in replication
      2. No information in original paper
         1. Ought to be included
2. Where do the differences in performance come from?
   1. Methods (as will be discussed in next section)

### 4.1.2 Methods (comparison)

#### 4.1.2.1 Data

* Language/nationality
  + Biased because of difference in labeling
    - This: Danish diagnostics
    - Original: Chinese, Malay, Indian diagnostics
  + Biased because of difference in language
    - This: Danish
    - Original: 3 Countries, with different languages
* Task
  + This: mid-level difficulty; description of triangles. No social component
  + Original: high-level difficulty; interview. Social component
* Data quantity
  + This: More participants with shorter recordings
  + Original: Fewer participants with longer recordings
* Sound quality
  + This: Difference in recording equipment
  + Original: Maybe?
* What contributed to the differences in performance? (If any)
  + Possibly all. Likely not sound quality to a large extent

#### 4.1.2.2 Feature selection

* Type of feature selection
  + This: LASSO - 5 fold
  + Original: PCA
  + Hard to replicate, given the sparse information on how PCA was used
    - Their feature selection method hard to follow
    - Could have been understood in two different ways
  + Specific feature selection method shouldn’t have a large impact on performance

#### 4.1.2.3 Machine learning algorithm

* Predicting (single participants, or same participants multiple times)
  + This: Predicting .wav files (several for each participant)
  + Original: Predicting participants
  + Does this matter?
* Ensemble modeling vs. Single machine learning algorithm
  + Stacking ensemble modeling
    - Better (if models are diverse, and generally good)
    - Only very slightly better
  + Single machine learning algorithm
    - Slightly worse

## 4.2 Pipeline

### 4.2.1 How did an implementation of pipeline in this replication work out?

1. Specifics:
   1. Data acquisition
      1. What did we do and why?
      2. Pros + cons + alternatives?
   2. Preprocessing
      1. What did we do and why?
      2. Pros + cons + alternatives?
   3. Data partitioning
      1. What did we do and why?
      2. Pros + cons + alternatives?
   4. Feature scaling
      1. What did we do and why?
      2. Pros + cons + alternatives?
   5. Feature selection
      1. What did we do and why?
      2. Pros + cons + alternatives?
   6. Model tuning (training, tuning and testing cycle)
      1. What did we do and why?
      2. Pros + cons + alternatives?
   7. Validation (and evaluation)
      1. What did we do and why?
      2. Pros + cons + alternatives?
   8. (Reflection + proper documentation)
2. Overall:
   1. Replication
      1. Possible but feature selection
         1. (Methods explained in condensed manner)
   2. Comparison of evaluation
      1. Good, but somewhat deficient
         1. More information on sexes and nationalities needed
   3. Replication got similar results
      1. Slightly different
      2. Differences in performance – where does it come from?
         1. Biased labels
         2. Difference in language
         3. Task differences
         4. Difference in algorithms
         5. Arbitrary choices for tuning
         6. A mixture (which mixture?) of all the above
      3. Some things might balance each other’s out, some might not
   4. All in all, fine! But are there still issues?

### 4.2.2 Insights on problems in research derived from doing a conservative replication

1. Curious that other studies have found much(!) higher accuracies
   1. Study 1
   2. Study 2
   3. Overfitting?
      1. My predictions on training 90% accuracy
      2. Scaling
2. Hard to know where differences in performance come from
   1. (All the differences on task, data, language, labeling etc.)
   2. Solution: More documentation on this and more reproductions
3. Bad documentation is insufficient for facilitating replication
   1. From practical experience
4. It is up to individual researchers and their experience to produce original studies and replications alike (not good)
   1. Arbitrary choices and handycrafts
      1. Tuning (C-parameters)
      2. Model type
      3. Paper (How do we choose defaults)
   2. From practical experience – not possible to find established pipeline and solutions

## 4.3 Further research

### 4.3.1 Taking everything into account: How does the pipeline help further research? (Wrap-up)

1. Meta – so these were the issues? What to do about it?
2. This pipeline DOES try to provide answers by:
   1. Avoiding in overfitting (as mentioned previously)
   2. Making it easier to compare results (as mentioned previously)
      1. Within or across sexes and nationalities (as mentioned previously)
   3. Making it easier to replicate (as mentioned previously)
   4. Enabling research to know locate the origin of differences in results (as mentioned previously)
      1. Biased labels
      2. Difference in language
      3. Task differences
      4. Difference in algorithms
      5. Arbitrary choices for tuning
      6. A mixture (which mixture?) of all the above
      7. Shedding light on arbitrary choices by providing information on it in the papers
3. This pipeline DOESN’T (alone) provide answers to:
   1. Too general and vague
      1. Doesn’t specify specifics -> very possible to do bad research
   2. Which factors apart from bad methods contribute to different ML results
      1. Answer ->
         1. Enough replications and research within each group might.
   3. Sharing of data and specific models (testing the same exact models on different data, not just method)
      1. Could also shed light on differences in language/biased labeling (diagnosistics)
4. In general, we need:
   1. More replications and research (using pipeline)
   2. A generally more open-science based approach

# 5. Conclusion

s

# 6. References

Abell, F., Happé, F., & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development*, *15*(1), 1–16. https://doi.org/10.1016/S0885-2014(00)00014-9

Alberto, P., Arndis, S., Vibeke, B., & Riccardo, F. (2019). *Voice Patterns in Schizophrenia: A systematic Review and Bayesian Meta-Analysis* [Preprint]. Bioinformatics. https://doi.org/10.1101/583815

Beck, K. I., Simonsen, A., Wang, H., Yang, L., Zhou, Y., & Bliksted, V. (2020). Cross-cultural comparison of theory of mind deficits in patients with schizophrenia from China and Denmark: Different aspects of ToM show different results. *Nordic Journal of Psychiatry*, 1–8.

Bliksted, V., Fagerlund, B., Weed, E., Frith, C., & Videbech, P. (2014). Social cognition and neurocognitive deficits in first-episode schizophrenia. *Schizophrenia Research*, *153*(1), 9–17. https://doi.org/10.1016/j.schres.2014.01.010

Bliksted, V., Frith, C., Videbech, P., Fagerlund, B., Emborg, C., Simonsen, A., Roepstorff, A., & Campbell-Meiklejohn, D. (2019). Hyper-and hypomentalizing in patients with first-episode schizophrenia: FMRI and behavioral studies. *Schizophrenia Bulletin*, *45*(2), 377–385.

Chakraborty, D., Yang, Z., Tahir, Y., Maszczyk, T., Dauwels, J., Thalmann, N., Zheng, J., Maniam, Y., Amirah, N., & Tan, B. L. (2018). Prediction of negative symptoms of schizophrenia from emotion related low-level speech signals. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6024–6028.

FabFilter Software Instruments. (2018). *FabFilter* (Fabfilter pro-q 2.) [Computer software].

Friedman, J., Hastie, T., & Tibshirani, R. (2010). *Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software*. *33(1)*, 1–22.

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, *101*(46), 16385–16389. https://doi.org/10.1073/pnas.0403723101

iZotope Inc. (2018). *IZotope RX 6*.

Olsen, L. (2018). *Automatically diagnosing mental disorders from voice: A deep learning approach*.

Olsen, L. (2020). *groupdata2: Creating Groups from Data* (1.3.0) [Computer software]. https://CRAN.R-project.org/package=groupdata2

*Statistics Denmark*. (n.d.). Retrieved 11 November 2020, from https://www.dst.dk/en/

Zivetz, L. (1992). *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines* (Vol. 1). World Health Organization.

# 8. Appendix

s