

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](#_Toc58320570)

[1.1 Schizophrenia and biomarkers 2](#_Toc58320571)

[1.1.1 Schizophrenia and voice atypicalities 2](#_Toc58320572)

[1.2 Machine learning for detection of acoustic patterns 3](#_Toc58320573)

[1.2.1 Prospects of machine learning in classifying schizophrenia 3](#_Toc58320574)

[1.2.2 Current limitations in the literature 3](#_Toc58320575)

[1.3 Alleviating current limitations 3](#_Toc58320576)

[1.3.1 Through replications and conservative ML implementation 3](#_Toc58320577)

[1.3.2 A rigorous ML pipeline 3](#_Toc58320578)

[1.3.3 Purpose of paper 4](#_Toc58320579)

[2. Methods 4](#_Toc58320580)

[2.1 Pipeline and the methods section summarized 4](#_Toc58320581)

[2.2 Literature search and choice of replication 6](#_Toc58320582)

[2.3 Data 6](#_Toc58320583)

[2.3.1 Data sources 6](#_Toc58320584)

[2.3.2 Participants 7](#_Toc58320585)

[2.3.2 Procedure/task 7](#_Toc58320586)

[2.4 Preprocessing 8](#_Toc58320587)

[2.4.1 Cleaning of audio files 8](#_Toc58320588)

[2.4.2 Feature extraction from audio files 8](#_Toc58320589)

[2.5 Partitioning 8](#_Toc58320590)

[2.6 Normalization 9](#_Toc58320591)

[2.5 Feature selection using LASSO 9](#_Toc58320592)

[2.5.1 Motivation for using LASSO 9](#_Toc58320593)

[2.5.2 What is L2 regularization? 10](#_Toc58320594)

[2.5.4 Feature selection 11](#_Toc58320595)

[2.6 Model training, testing and parameter tuning 12](#_Toc58320596)

[2.7 Evaluation metrics 12](#_Toc58320597)

[3. Results 13](#_Toc58320598)

[4. Discussion 1](#_Toc58320599)

[4.1 Results and comparison between original study and replication 1](#_Toc58320600)

[4.1.1 Performance comparison to original study 1](#_Toc58320601)

[4.1.2 Methods comparison to original study (where were the differences?) 2](#_Toc58320602)

[4.2 Pipeline 3](#_Toc58320603)

[4.2.1 How did an implementation of pipeline in this replication work out? (specific level – each pipeline step: pros + cons + alternatives) 3](#_Toc58320604)

[4.2.2 How did an implementation of pipeline in this replication work out? (broad/general level – could the replication be carried out? is it useful?) 3](#_Toc58320605)

[4.3 Further research 4](#_Toc58320606)

[4.3.1 Insights on general problems in research (knowledge gained from doing a conservative replication) 4](#_Toc58320607)

[4.3.2 Benefits and limitations of the use of this pipeline in further research (Wrap-up) 4](#_Toc58320608)

[5. Conclusion (might be scrapped) 5](#_Toc58320609)

[6. References 5](#_Toc58320610)

[7. Appendix 6](#_Toc58320611)

**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 classifying with voice in conjunction with other things)
         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. No information on performance across languages/sex/how and who are diagnosed (across countries)

## 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
      3. Evaluate implementation

# 2. Methods

## 2.1 Pipeline and the methods section summarized

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 specifically replicate the paper by Chakraborty et al. from 2018 (Chakraborty et al., 2018). Additionally, a short summary will be provided along with two figures in order to provide an overview of the process and showcase how it followed our proposed pipeline. One figure 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 green 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 in this replication. **3) Partitioning.** The data was partitioned into a training and 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 folds were used in 5 different splits – with each split having a training set consisting of 4 out of the 5 folds and a testing set consisting of the remaining fold. All training sets had their features L2 regularized; resulting in a feature sets for each split. These feature sets only contained the most relevant features for the classification of schizophrenia patients.   
This process did in other words produce 5 splits of training and testing data. It also produced 5 feature sets. Each of these feature sets had been selected on the basis of 4/5th of the full training data, which meant that the remaining 1/5th of the training data (the testing set) 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) using the respective feature sets. The fit models were then tested on the matching test set (last 1/5th). The predictions were then evaluated based on their classification performance and C and Gamma parameters were tuned in the models. After tuning, the models were then tested again – repeating this process until need for a satisfactory performance level was met.

**9) Validation on holdout set.** Finally, the 5 models were tested on the holdout set. An ensemble model was also constructed. This model also predicted the holdout data, but by using the majority vote of the other 5 models as its prediction. Performance on the holdout set was then evaluated for the 6 models with the use of relevant metrics. Performance of the sexes was also calculated separately 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 2014 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 the healthy 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 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 done using the package groupdata2 and was carried out semi-randomly (Olsen, 2020). The partitioning kept each participant ID only within either the resulting training set or the resulting 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 in 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 holdout set. The constituents of the holdout set were the feature vectors 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; calculating performance of females with predictions on a single female would either yield a 100% or 0% accuracy, neither of which would necessarily be telling for a models true performance.

## 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 of both sets used the min. and the max. value for each feature, only from the training 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. This procedure 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. ElasticNet 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 for all 5 feature sets 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 for each time it is performed. Those that have not been shrunken to zero are selected as relevant features for predicting patients from controls. For a visualization of lambda misclassification plot, 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 all 5 training splits, resulting in a feature set for each of them (see appendix x\* for list of these feature sets). An illustration of the feature selection for a single split (split 1), can be seen below in figure 4.

Graphical user interface

Description automatically generated

Figure 4:

*Figure showing the process of feature selection on train 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 accumulated errors from all relevant lambda values have been obtained.*

*This entire procedure is then repeated for each of the remaining 4 training 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 1 was tested on test set 1, etc. Performance was then 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 was 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 precision and recall 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 the two ratios and gives an overall understanding of the classification performance for each class. By calculating the arithmetic mean of the two F1-scores (harmonic means) for a model, the single score of macro-F1 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 all 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. The latter uses the abbreviations TD and SZ which mean ‘typically developed’ and ‘schizophrenia”, respectively.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 row 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.*

|  |  |  |  |
| --- | --- | --- | --- |
| 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

This section will first compare the results of this replication with the results of the original paper. Moreover, the difference in how the ML study was carried out will be discussed and the differences’ influence on performance is reflected upon.

Secondly the both the general pipeline as well as the specific implementation of the pipeline in this replication will be evaluated. The question “*How did an implementation of the proposed pipeline in this replication work out?*“ will be addressed. This will be done on two levels:  
1) on the level of this specific replication (evaluating the choices for the 8 steps) and   
2) on the macro-level (was it possible to replicate using the general pipeline? Is it useful?)

Finally, the general problems within this research area will be addressed using the insights from this replication. The use of the proposed general pipeline in future research will also be discussed, looking into both benefits and limitations. On this basis, the prospects for conducting further research are then assessed.

## 4.1 Results and comparison between original study and replication

### 4.1.1 Performance comparison to original study

1. Performance of models on test
   1. Evaluation metrics (F1, accuracy, precision + recall,)
      1. F1-score for model overall
      2. F1-scores for patients and controls respectively
      3. Accuracy (not very telling)
      4. Precision + Recall
   2. 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)

This section will compare the original papers performance with the performance of the ensemble model on the holdout set from this study.

The reason for looking at the performance when predicting the holdout set is because it gives information about the out-of-sample capabilities of the model. This contrasts with predicting the training data, as this would give no idea of the generalizability of the model.   
The reason for looking at the performance of the ensemble model instead of for example the best performing model, is because it is likely to be the most robust model. The fact that some of the training sets and their appertaining feature sets produced better predictions on this specific holdout set (e.g. train 4), is likely only due to random chance. The holdout dataset will undoubtedly have more in common with some of the training splits than others. The ensemble model will be more robust and generalizable since it simply is based on more data. Larger samples will more closely approximate the true population \* cite \*.

The ensemble model achieved an overall accuracy of 70.32%, which is lower than the original paper’s 70.49%. This can be misleading however, as it does not account for differences in baseline accuracy – the original study employed a testing set of 66.67% patients, while this replication had majority group of 51.87%.  
The macro average F1-score gives a better measure of performance. The original paper had an F1-score of 0.77 – higher than the F1-score of 0.703 in this replication. When looking at the isolated F1-scores for classifying patients and controls it can be seen that both the ensemble model and the model of the original study classified controls equally well. The original model did, however, achieve a higher F1-score when classifying patients. The F1-score alone doesn’t enquire into the ratio between precision and recall – the models could have a relatively high recall and low precision or vice versa. This was, however, not the case. The recall and precision rates were evenly balanced for both models.

SEX - KOMMET HERTIL - !!!!!!!!!!!!!!!!

Better overall accuracy in original (70.32% vs 79.49%)

Replication better improvement to baseline predictions (12.82 percentage point increase, 19.2290% increase in accuracy vs 18.45 percentage point increase, 35.5696% increase in accuracy)

However, baseline is different as original study employed a holdout set of 2/3rd patients – so not even the percentage point increase is telling.

F1-scores better measure

Better macro avg. F1-score in original (0.703 vs. 0.77).

However, F1-score for TD is the same across studies -> the models are equally good at classifying controls correctly.

F1-score for SZ is higher in original. Higher harmonic mean of recall and precision within SZ predictions.

F1-score alone doesn’t enquire into the rate between precision and recall – was there a relatively high recall and low precision? Or were they equally balanced?

original SZ and TD recall and precision reasonably balanced, same for replication

given male and female physiological differences, voice differs. Reasonable to assume this could have an impact on the classifier – also even though both datasets were roughly equally balanced on this matter. F1-scores between male and female roughly the same, also for both sz and td (male sz= 0.714, male td = 0.698, female sz = 0.697, female td = 0.703).

no information given on the original study by Chakraborty but given the even balancing their model might very well also predict equally well for both sexes. This ought to have been included, though – it isn’t necessarily the case.

Replication

SZ F1-score = 0.706

TD F1-score = 0.700

Macro avg. F1-score = 0.703

Accuracy = 70.32%

Baseline accuracy = 51.87%

Balanced accuracy = 70.45 %

SZ recall = 0.739

SZ precision = 0.675

TD recall = 0.670

TD precision = 0.734

Original

SZ F1-score = 0.84

TD F1-score = 0.70

Macro avg. F1-score = 0.77

Accuracy = 79.49 %

Baseline accuracy = 66.67 %

SZ recall = 0.83

SZ precision = 0.86

TD recall = 0.73

TD precision = 0.68

### 4.1.2 Methods comparison to original study (where were the differences?)

#### 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 Partitioning

* There is none in the original. They use the same data for both training and testing.
* Training set
  + This: Mostly balanced on gender. Mostly balanced on diagnosis
  + Original: Balanced on gender. Very unbalanced on diagnosis
* Holdout set
  + This: Very balanced on gender. Very balanced on diagnosis
  + Original: Balanced on gender. Very unbalanced on diagnosis

#### 4.1.2.3 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

#### 4.1.2.4 Machine learning algorithm

* Predicting (single participants, or same participants multiple times)
  + This: Predicting .wav files (several for each participant)
  + Original: Predicting participants
  + Should not have large impact on performance
* 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
  + Should not have large impact on performance

## 4.2 Pipeline

### 4.2.1 How did an implementation of pipeline in this replication work out? (specific level – each pipeline step: pros + cons + alternatives)

1. Data acquisition
   1. What did we do and why? Pros + cons + alternatives?
2. Preprocessing
   1. What did we do and why? Pros + cons + alternatives?
3. Data partitioning
   1. What did we do and why? Pros + cons + alternatives?
4. Feature scaling
   1. What did we do and why? Pros + cons + alternatives?
5. Feature selection
   1. What did we do and why? Pros + cons + alternatives?
6. Model tuning (training, tuning and testing cycle)
   1. What did we do and why? Pros + cons + alternatives?
7. Validation (and evaluation)
   1. What did we do and why? Pros + cons + alternatives?
8. Reflection + proper documentation
   1. Did we do this and why?

### 4.2.2 How did an implementation of pipeline in this replication work out? (broad/general level – could the replication be carried out? is it useful?)

1. Replication
   1. Possible but feature selection not so much
      1. (Methods explained in condensed manner in original)
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. Slight difference in performance – where 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. Reflection + proper documentation
   1. Yes – remembered to do this
5. Wrap up – Replication seems to have worked out OK

## 4.3 Further research

### 4.3.1 Insights on general problems in research (knowledge gained from doing a conservative replication)

1. Curious that other studies have found much(!) higher accuracies
   1. Study 1 with much higher accuracy
   2. Study 2 with much higher accuracy
   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 to narrow down.
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

### 4.3.2 Benefits and limitations of the use of this pipeline in further research and going forward (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 (might be scrapped)

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.

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.

# 7. Appendix

s