

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

[Abstract 3](#_Toc57284910)

[1. Introduction 3](#_Toc57284911)

[1.1 Schizophrenia and biomarkers 3](#_Toc57284912)

[1.1.1 Schizophrenia 3](#_Toc57284913)

[1.1.2 Biomarkers and voice atypicalities 3](#_Toc57284914)

[1.2 Machine learning for detection of acoustic patterns 5](#_Toc57284915)

[1.2.1 Prospects of machine learning in classifying schizophrenia 5](#_Toc57284916)

[1.2.2 Current limitations in the literature 7](#_Toc57284917)

[1.3 Alleviating current limitations 7](#_Toc57284918)

[1.3.1 Through replications 7](#_Toc57284919)

[1.3.2 Through proper ML implementation 7](#_Toc57284920)

[1.3.3 Thesis statement / purpose of paper 9](#_Toc57284921)

[2. Materials and Methods 10](#_Toc57284922)

[2.1 Pipeline 10](#_Toc57284923)

[2.2 Literature search and choice of replication 12](#_Toc57284924)

[2.3 Data 12](#_Toc57284925)

[2.3.1 Data sources 12](#_Toc57284926)

[2.3.2 Participants 12](#_Toc57284927)

[2.3.2 Procedure/task 13](#_Toc57284928)

[2.4 Preprocessing 14](#_Toc57284929)

[2.4.1 Cleaning of audio files 14](#_Toc57284930)

[2.4.2 Feature extraction from audio files 14](#_Toc57284931)

[2.4.3 Partitioning 14](#_Toc57284932)

[2.4.4 Normalization 15](#_Toc57284933)

[2.5 Feature selection using LASSO 15](#_Toc57284934)

[2.5.1 Motivation for using LASSO 15](#_Toc57284935)

[2.5.2 What is L2 regularization? 16](#_Toc57284936)

[2.5.4 Feature selection 17](#_Toc57284937)

[2.6 Model and model tuning 18](#_Toc57284938)

[2.7 Evaluation metrics 18](#_Toc57284939)

[3. Results 19](#_Toc57284940)

[4. Discussion 2](#_Toc57284941)

[4.1 Results and replication comparison 2](#_Toc57284942)

[4.1.1 Performance 2](#_Toc57284943)

[4.1.2 Data 2](#_Toc57284944)

[4.1.3 Feature selection 3](#_Toc57284945)

[4.1.4 Methods () 4](#_Toc57284946)

[4.2 Pipeline 4](#_Toc57284947)

[4.2.1 Implementation of pipeline in replication 4](#_Toc57284948)

[4.2.2 Pointers about ML within this field, from the implementation of the pipeline 5](#_Toc57284949)

[4.3 Where does research go from here? 5](#_Toc57284950)

[4.3.1 Need for strict research principles 5](#_Toc57284951)

[4.3.2 More replications 6](#_Toc57284952)

[4.3.3 Better transparency, descriptions, + more open sciency 6](#_Toc57284953)

[5. Conclusion 6](#_Toc57284954)

[6. References 6](#_Toc57284955)

# Abstract

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

# 1. Introduction

## 1.1 Schizophrenia and biomarkers

### 1.1.1 Schizophrenia

s

### 1.1.2 Biomarkers and voice atypicalities

s

## 1.2 Machine learning for detection of acoustic patterns

### 1.2.1 Prospects of machine learning in classifying schizophrenia

s

### 1.2.2 Current limitations in the literature

s

## 1.3 Alleviating current limitations

### 1.3.1 Through replications

s

### 1.3.2 Through proper ML implementation

s

### 1.3.3 Thesis statement / purpose of paper

s

# 2. Materials and methods

## 2.1 Pipeline

How this replication follows pipeline (short overview of methods section)

The replication of this paper follows and provides an exemplification of the steps the pipeline consists of (fig. 1). 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 a long with a figure which visualizes the process (fig. 2).

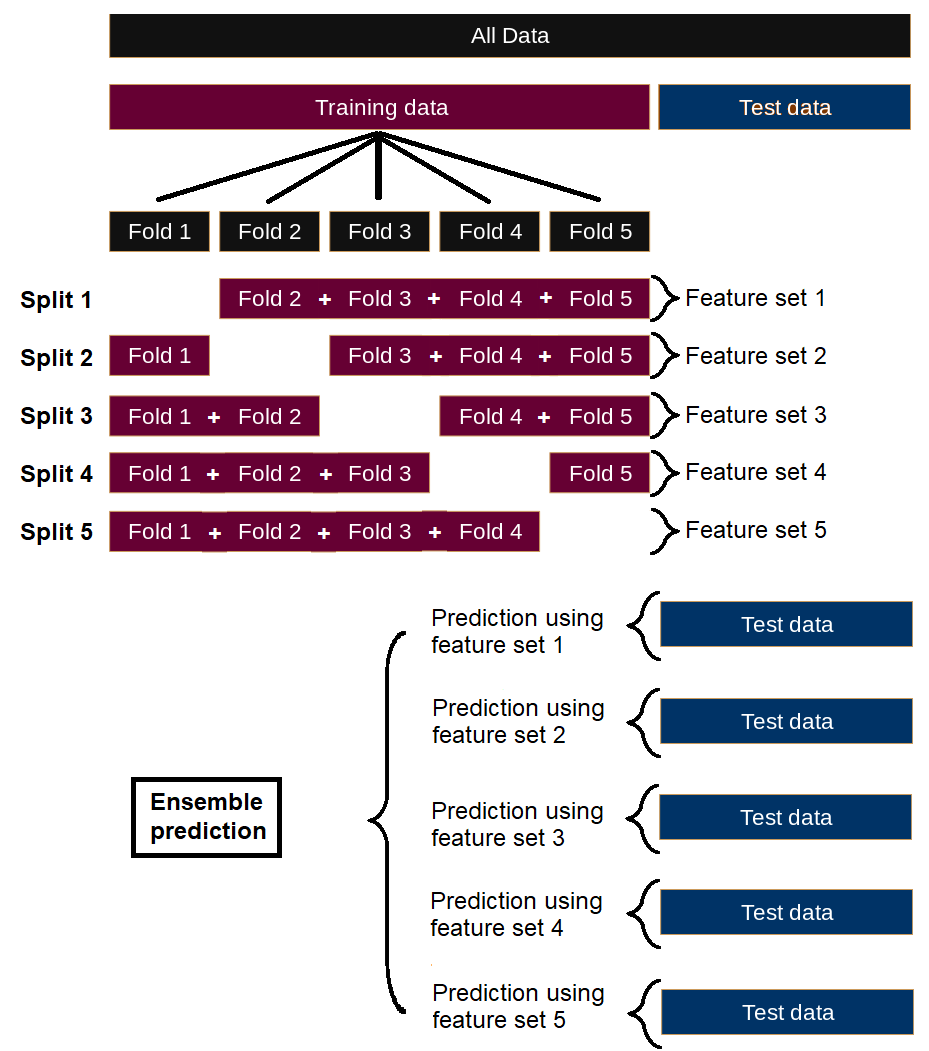


Fig. x \*.

*An overview visualization of this specific machine learning process - partitioning, feature reduction sets and predictions of the holdout set*

First and foremost, cleaned voice data was acquired from which relevant features were extracted (the feature list was in this replication determined by the original study). The voice data was then partitioned into a training and a holdout set in an 80/20 partitioning process that kept IDs separate in each partition. Feature scaling was then performed using min/max normalization such that it did not allow data leakage from the training to the holdout set. As an ensemble model was desired, we then performed feature selection using LASSO on 5 different splits of the training data. The 5 resulting feature lists where then used as parameters for training 5 distinct SVM linear kernel classification models. These model in turn had their parameters tuned before being tested on a test portion of the training data that the models had not seen. After tuning the parameters until a desired outcome was found, the training data was used to train 5 models, each with its own feature set and subsequently predicted the holdout set. The majority vote of a given holdout voice file was then recorded and this constituted the ensemble models predictions. Finally, all models were evaluated based on their performance, with appropriate metrics.

## 2.2 Literature search and choice of replication

A literature search for papers, dissertations and unpublished manuscripts was conducted for finding the 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 on after taking these factors into consideration (Chakraborty et al., 2018).

\* Expand on this? \*

## 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 (FHA) (Abell et al., 2000). All participant went through 8 such trials, except for in the study from 2015 by Bliksted et al., where the participants were presented with 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 x \*.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 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 x \* :

*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

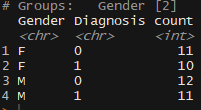
regression coefficients and linear and quadratic error, standard deviation, skewness, kurtosis,

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.

### 2.4.3 Partitioning

To be able to evaluate the performance of the model the dataset was partitioned into a training set and a test 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 and test sets. This prevented leakage of information from the training set to the test set, which otherwise would have led to overfitting and finally unprecise values for the evaluation. Moreover, to avoid a skewed distribution of sex or diagnosis between sets (e.g. ending up with only males/controls in the test set as a result of a random partitioning), sex and controls/patients were evenly distributed in the partitioning.



### 2.4.4 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:  

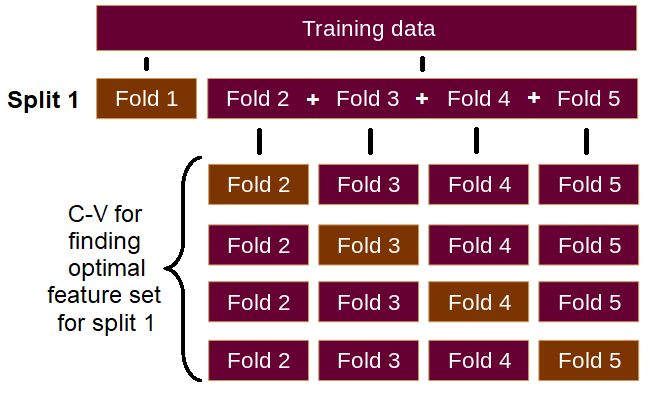

Since this method requires a lambda value (λ), the optimal lambda value also had to be found. The lambda value producing the minimum value in the loss function (lambda.min) was first computed. This was done by testing a range of lambda values using Leave-One-Out CV\*\*\*\* 5-FOLD? SEE FIGURES \*\*\*\*\*\*\* (LOO-CV). 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 fig. x \*.



Fig. x \* :

*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.   
Figure x \* :

*Figure showing the process of feature selection for one of the splits:*

*The training data is divided up into 5 folds. One fold is then excluded. 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 and model tuning

s

## 2.7 Evaluation metrics

# 3. Results

s

# 4. Discussion

## 4.1 Results and replication comparison

### 4.1.1 Performance

s

### 4.1.2 Data

t

### 4.1.3 Feature selection

s

### 4.1.4 Methods ()

s

## 4.2 Pipeline

### 4.2.1 Implementation of pipeline in replication

s

### 4.2.2 Pointers about ML within this field, from the implementation of the pipeline

s

## 4.3 Where does research go from here?

### 4.3.1 Need for strict research principles

s

### 4.3.2 More replications

s

### 4.3.3 Better transparency, descriptions, + more open sciency

s

# 5. Conclusion

s

# 6. Acknowledgements

s

# 7. 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