

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](#_Toc59200878)

[1.1 Schizophrenia and voice as a biomarker 2](#_Toc59200879)

[1.2 Machine learning of acoustic patterns for detection of schizophrenia 3](#_Toc59200880)

[1.2.1 Prospects of machine learning in classifying schizophrenia 3](#_Toc59200881)

[1.2.2 Current limitations in the literature 4](#_Toc59200882)

[1.3 Alleviating current limitations 5](#_Toc59200883)

[1.3.1 Through replications and conservative ML implementation 5](#_Toc59200884)

[1.3.2 A rigorous ML pipeline 5](#_Toc59200885)

[1.3.3 Purpose of paper 8](#_Toc59200886)

[2. Methods 9](#_Toc59200887)

[2.1 Pipeline implementation – an overview of the methods 9](#_Toc59200888)

[2.2 Literature search for choice of replication 11](#_Toc59200889)

[2.3 Data 11](#_Toc59200890)

[2.3.1 Data sources 11](#_Toc59200891)

[2.3.2 Participants 12](#_Toc59200892)

[2.3.2 Procedure/task 12](#_Toc59200893)

[2.4 Preprocessing 13](#_Toc59200894)

[2.4.1 Cleaning of audio files 13](#_Toc59200895)

[2.4.2 Feature extraction from audio files 13](#_Toc59200896)

[2.5 Partitioning 14](#_Toc59200897)

[2.6 Normalization 14](#_Toc59200898)

[2.7 Feature selection 14](#_Toc59200899)

[2.7.1 L2 regularization 14](#_Toc59200900)

[2.7.2 Feature selection 16](#_Toc59200901)

[2.8 Model training, testing and parameter tuning 17](#_Toc59200902)

[2.9 Evaluation metrics 18](#_Toc59200903)

[2.10 Differences between replication and original study 18](#_Toc59200904)

[3. Results 19](#_Toc59200905)

[4. Discussion 21](#_Toc59200906)

[4.1 Performance results and comparison of performance to original study 21](#_Toc59200907)

[4.2 Evaluating specific pipeline implementation (and discussing differences from original study) 22](#_Toc59200908)

[4.3 Limitations and prospects of the proposed overall pipeline 25](#_Toc59200909)

[5. References 26](#_Toc59200910)

[6. Appendix 35](#_Toc59200911)

[6.1 Relevant studies 35](#_Toc59200912)

[6.2 Feature lists after L2 Regularization 36](#_Toc59200913)

Abstract

Can machine learning (ML) applied to voice data be used as a tool to help diagnose people with schizophrenia? Numerous studies have shown high accuracies when classifying schizophrenia, but results are widely heterogenous, as concluded in the latest meta study within the field (Parola et al., 2019). Little work has investigated the validity and robustness of the previous findings, and few replications shed light on the generalizability of the studies. Currently there is no consensus on which machine learning pipeline should be applied for optimal classification of schizophrenic patients.

This study provides a proposal for a general machine learning pipeline suitable for future research into this topic, along with the information necessary for implementing both rigorous and conservative machine learning models. As the pipeline is meant to be broad and general, an example of its more specific implementation is also provided for clarifying purposes. It is exemplified through a replication of the promising study by Chakraborty et al (Chakraborty et al., 2018), and thus also has the benefit of contributing to science within this area that is largely deprived of replications. A macro average F1-score of 0.70 was found – notably lower than the original study’s 0.77. As the replication employed a dissimilar dataset and slightly diverging methods, these differences were discussed in relation to the results. Subsequently, the proposed overall pipeline was scrutinized and given its limitations, further development on the pipeline was suggested. Finally, this paper advocates for a more open and cumulative scientific community.

Keywords: schizophrenia, speech signal, acoustic features, biomarker, machine learning

# 1. Introduction

## 1.1 Schizophrenia and voice as a biomarker

Schizophrenia has been associated with several language and voice impairments. The qualitative described symptoms of alogia and blunted affect have been found in large parts of the population suffering from schizophrenia (Andreasen et al., 1995; Cohen et al., 2012). Alogia refers to characteristics such as poverty of speech, latency of speech, blocking, while blunted affect refers to a decrease in emotional expression and a lack of vocal intonation. Schizophrenia is furthermore associated with various other differences that range from higher-order semantic language impairments and semantic processing (Covington et al., 2005; Kuperberg, 2010) to differences in low-level acoustic signals such as shimmer and jitter (Kliper et al., 2016). A recent meta-analytic study from 2019 by Alberto et al., has systematically reviewed the accumulated evidence for distinctive acoustic patterns in schizophrenia (Parola et al., 2019). They found modest effects for proportion of spoken time, speech rate, pauses, and pitch variability, while pause duration proved to be a relatively strong predictor.

The language and speech disturbances are used in the clinical assessment process and proven helpful for identifying those individuals that are at a high risk for developing psychosis – even before onset (Bearden et al., 2011; DeVylder et al., 2014; Sichlinger et al., 2019). They have furthermore allowed for tracking psychotic symptoms and predicting progression in symptoms (Bearden et al., 2011; Corcoran et al., 2020; Morice & Ingram, 1983; Solomon et al., 2011). There is, however, a big drawback to the current use of speech in schizophrenia. Speech is being manually annotated or rated by expert raters, which is time extensive and requires training of the raters. This makes the procedure expensive and impractical on a large scale. Moreover, there is a chance that only the most extreme cases are picked up on, when using these manual assessments (Hitczenko et al., 2020). As the prospects of using speech clinically are ample but merely impractical on a larger scale, recent endeavors have been made to automate this method using supervised machine learning (ML) approaches.

## 1.2 Machine learning of acoustic patterns for detection of schizophrenia

### 1.2.1 Prospects of machine learning in classifying schizophrenia

Supervised ML classification works by learning patterns in some data set and can then subsequently be used to predict, using the learned patterns. The ‘learning’ part practically means building a model of the distribution of class labels (e.g. schizophrenic/non-schizophrenic) from predictor variables (e.g. acoustic features from speech). The ‘machine’ entails that the process is automated, which allows for finding complex, multivariate and sometimes non-linear relationships between multiple features in conjunction and class labels (Kotsiantis et al., 2007). After training a model, it can then be used to assign class labels to the testing instances where the class labels are unknown, but where the predictor values are known.

ML has the potential of supporting clinical evaluations. While the currently implemented manual assessment is impractical on a large scale, ML is not. It would allow for preemptively identifying those at risk for developing schizophrenia on a large scale and give clinicians an effortless way of tracking and predicting progressions in symptoms. Furthermore, judgements are objective given their automated nature (Hitczenko et al., 2020). Classification algorithms have been able to classify schizophrenia with accuracies between 70% and 95% (Martínez-Sánchez et al., 2015; Rapcan et al., 2010; Stassen et al., 1995; Tahir et al., 2019). If computational methods can achieve these rates of correct predictions, they may very well be applied clinically.

### 1.2.2 Limitations of machine learning methods in schizophrenia

Although the method of machine learning looks promising at first glance, some substantial hurdles are still standing in the way before computational methods can be instantiated clinically.

One hurdle is the overfitting issues that exists within the field (Vabalas et al., 2019; Voleti et al., 2019). Overfitting is the term for having models learn and rely on spurious correlations between features (acoustic features within this field) and a class (such as a diagnosis). Studies with overfit models might publish good performance, but the models have low generalizability and would predict poorly new data (Dietterich, 1995). This is likely to be the cause of the large discrepancy in results across studies, that have been found (Hitczenko et al., 2020).

Another hurdle within the literature is the diversity in ways of conducting research. As this field of research is relatively new, no universally accepted way of conducting ML exists. As a result, studies vary considerably in methods, method quality, transparency and documentation. Not only does this make it hard to compare studies, but it also makes it difficult to pinpoint which methods, feature sets, or datatypes produces the best results. When a study finds a classification rate of 87.5% (Martínez-Sánchez et al., 2015), while another finds a rate of 79.5% (Chakraborty et al., 2018) can the reason for the difference in performance be investigated? The difference in performance might be due to one study using LDA classification as a method, while the other uses SVM. It could also be due to one study using features related to emotion as predictors, while the other does not. Furthermore, studies also vary in their way of documenting both methods and results. If the methods or the results are inadequate for comparing across studies, then it is hard to pinpoint which factors cause what (Hitczenko et al., 2020).

## 1.3 Alleviating current limitations

### 1.3.1 Through replications and conservative ML implementation

Replications and studies differing slightly from past work (such as on nationality of participants) give clear insights into the impact of specific factors (e.g. showing that cross-cultural differences impact results). Proper ML implementation ensures that these inferences can be made, as this ensures that studies are a) replicable – the studies must be transparent and properly document the entire process of conducting the study. b) using proper and conservative methods – results are only insightful if the models producing them do not suffer from problems of overfitting or bias. To alleviate limitations within this research area, we must ensure that the two previously mentioned criteria are met. But what constitutes a proper conservative ML implementation?

To try and provide the foundation for further research, this paper will outline what is required to perform proper conservative ML research. A rigorous pipeline will thus be provided, allowing for better ML practice and for improving the conditions for comparisons of results between studies.

### 1.3.2 A rigorous ML pipeline

A ML pipeline consists of several steps to train a model and operate workflow guidelines, from which predictive algorithms can be created. A pipeline can be used to support and streamline research, making comparison to other work easier (Guzzetta et al., 2010; Olson & Moore, 2016; Samad & Witherow, 2018). In turn, this will enable insights of the impact of specific methods, features or populations on machine learning within a given research field.

The pipeline that this paper is presenting is general and broad, with aspirations of being widely inclusive. However broad, it will narrow in the range of options to ensure that necessary requirements for good ML conduct are being met. The pipeline will be divided up into 9 steps. The steps will not specify exactly how they ought to be carried out, which means that remains crucial to scrutinize the methods chosen for a ML study. Proper and transparent documentation is crucial, as in all research, but perhaps especially within a field that suffers from little replicability. It is intended to be beneficial for ML attempts using voice as a predictor within the field of schizophrenia but may be directly applicable to various other fields such as in research on autism or depression.

**Data acquisition:** When gathering data for a machine learning analysis, it is important to know what data is being processed, to avoid pitfalls. Neither of following factors are going pose any issues if dealt with appropriately but might inflict difficulties if neglected.

First, it is important to be wary of any bias that might arise in the model as a result of subgroups or unintended structure in the participant pool from where the data comes from. Sociodemographic factors, such as educational level, age, race, sex have been known to cause a wide array of harmful bias across research fields (Blodgett et al., 2020; Hitczenko et al., 2020), but additional factors such as medication and severity of symptoms might also contribute to biases. Computational methods on voice in particular have had issues, with voice recognition software being shown to perform worse on the black population in the US (Koenecke et al., 2020) or that vocal expression was found not to be modulated by schizophrenia when accounting for sociodemographic background and task (Cohen et al., 2016).

Secondly, data quantity is important. Internal and external validity of a study have been found to undermined by small sample sizes (Faber & Fonseca, 2014), and there has been found an association between small sample sizes and biased performance in ML studies classifying diagnosis from voice (Vabalas et al., 2019). Another factor that has to be taken into account is the task from which the recordings are derived. Cognitive and social load has been found to increase the effects of schizophrenia in the acoustic signal (Parola et al., 2019). Finally, irrelevant recording identifiers must be controlled for. Background noise, room ambience or recording equipment and settings should ideally be uniform across diagnosis – having all speech from schizophrenics be recorded in one room, and the healthy controls in another could cause potential problems as the acoustic features of participants might be altered by room acoustics.

**Preprocessing:** Preprocessing includes noise removal and data augmentation. This step may either rid the recordings of unwanted signals, or apply additional signals to control for confounds – such as adding convolutional reverb to make the model more robust to differences in reverb coming from confounding factors such as the acoustic qualities of recording rooms (Olsen, 2018).Since raw recordings cannot be used to predict, features have to be extracted from the speech. The choice of features set can be driven by theory, by choices of past studies or can be entirely explorative.

**Data partitioning**: As train/holdout splits have found to be more robust and provide less balanced results, compared to K-fold cross-validation, it is recommended that the data is split into a training and a holdout set (Vabalas et al., 2019). The training set can furthermore be divided up into a training and a validation set, to benefit from better hyperparameter tuning as described under model tuning (Schratz et al., 2019).   
The ratio of train/test has an impact. A larger training allows for the model to better learn the patterns in the data, while a larger holdout set allows for a more accurate measure of performance - having three voice recordings to predict could only result in an accuracy of either 0%, 33.33%, 66.67% or 100% accuracy even given a true accuracy 70% \* cite \*. Although there is some basis for choosing the split, the optimal split size is considered to be rather arbitrary – there is no scientific consensus on what is optimal, although 70/30 or 80/20 is often used. \* cite \*.  
If one deals with an unbalanced dataset, whether it means having few schizophrenia or female instances, it is also important to take this into account during partitioning. An unbalanced training set of 4 male patients and 2 female controls, might simply lead to the model predicting ‘schizophrenic’ to all cases where the acoustic features are specific to males. Instead of learning the acoustic patterns of schizophrenia it would learn the acoustic patterns of males.  
An unbalanced holdout set consisting of too few members of a group – whether it be sex, nationality of diagnosis - has the problem previously mentioned problem of not allowing for an accurate measure of performance within that group. This performance measure within one group is the only way of seeing whether the model learned based on some confounding factor (such as sex – if for example all controls were female).

**Feature scaling:** Feature scaling is a necessary step for most algorithms to function properly. It has been known to improve performance, as well as decrease the computational load and avoid convergence issues \*cite \*. Many techniques for doing so exists and they have different merits. Regardless of technique, however, it is important to avoid scaling the pooled features from the training and holdout set – instead the scaling of both the training and holdout set should only use information (e.g. min-max values if using min-max normalization) from the training set. This ensures that no information can flow from the training to the holdout set, which otherwise would result in overfitting (Myrianthous, 2020).

**Feature selection:** It can be necessary to select a subset of features, if the extracted feature set contains many features. Feature selection is carried out in order to simplify the model and thereby reduce both complexity and computational power needed to run the model. Furthermore, feature selection has known to improve both predictive power and interpretability of the classifier. Features must only be selected on basis of information in the training set, and not on the pooled training and testing data before the partitioning. Selecting relevant features based on what is relevant in the holdout set is going to produce overfitting and low generalizability (Vabalas et al., 2019).   
Numerous feature selection techniques exist, and although choosing a technique might seem an arbitrary choice, it is not. Although in theory they perform the same task, in practice they do not perform equally well (Oreski et al., 2017). There simply is no silver bullet method, however, as the best individually performing feature selection technique depends on both dataset and classifier algorithm (Jović et al., 2015).

**Model training, tuning and testing:** Model training. Supervised machine learning covers a wide range of algorithms, that all produce models based on some set of training data that can be used for prediction. Common to most of them is the embedded use of hyperparameters - parameters with values that control the learning process of a given algorithm. Performance is critically dependent on hyperparameter settings and they must be specified before training a model (Hutter et al., 2014). However, determining the appropriate values can be complex (Claesen & De Moor, 2015). Some software implements automated ways of doing so, but at the present time, they do not necessarily determine the optimal values (Feurer & Hutter, 2019; Mantovani et al., 2016; Olson et al., 2017; Sanders & Giraud-Carrier, 2017; Thornton et al., 2013). But how can we tell if the model is tuned correctly?  
One of the benefits of partitioning the data up into a training, a validation and a holdout set is the possibility of validating the model on the validation set. After having the model trained on the training set with a given set of hyperparameters, see how well it predicts the validation set. Afterwards the hyperparameters can be tuned and a whole range of values might be tried out – the use of the validation set enables tracking of performance given different hyperparameter settings. This process may continue circularly until performance is satisfactory and by validating on the validation set, we can be assured that the model has not been overfit to the test set, thus making the test set suitable for as an evaluation of true performance.

**Validation:** Confusion matrices are of utmost importance when documenting performance of classifiers. They provide the complete picture of performance and all relevant metrics of performance can be calculated solely using the information of a confusion matrix. However, to make sense of the matrices, evaluation metrics ought to 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 can be provided for the whole model, 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). Since a model might have a high precision and a low recall rate (or vice versa), a F1-score used often used. 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 (one F1-score for the classification of patients, and one for controls), the single score of macro-F1 provides clear and quick overall insight into classification performance.   
Accuracy – the percentage of correct classifications - is said to give an intuitive impression of the performance and is regarded common practice. However, it can often be misleading (e.g. when evaluating performance on an unbalanced test set). 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.

### 1.3.3 Purpose of paper

To summarize; voice proves to be an important biomarker for schizophrenia with prospects of widespread application if automated.

Machine learning proves promising and appear to be able to distinguish and schizophrenia. However, the field of machine learning within this topic have issues with overfitting, lack of replications and comparability of results between studies as a result of large differences in methods between studies.

To alleviate these problems, this study provides a pipeline which assist in diminishing issues of overfitting as well as improving conditions for both replicability and comparison of results between studies.

As a way of providing an exemplification of the proposed general pipeline, as well as facilitate replications, this study will furthermore perform a replication of the ML study by Chakraborty and colleagues from 2018 (Chakraborty et al., 2018).

Finally, the implementation of the pipeline in the replication will be evaluated - and the potential limits and prospects of the pipeline will be discussed.

# 2. Methods

## 2.1 Pipeline implementation – an overview of the methods

The replication of this paper follows and provides an exemplification of the use of a rigorous pipeline - following the overall principles presented in 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 attempts to visualize the pipeline (figure 1) and one 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 implementation of the pipeline 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 988 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.** 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; meaning the feature estimates of the features that were less important were shrunk to zero, and thus essentially removed. This process resulted in a feature set with fewer than the original 988 features 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.

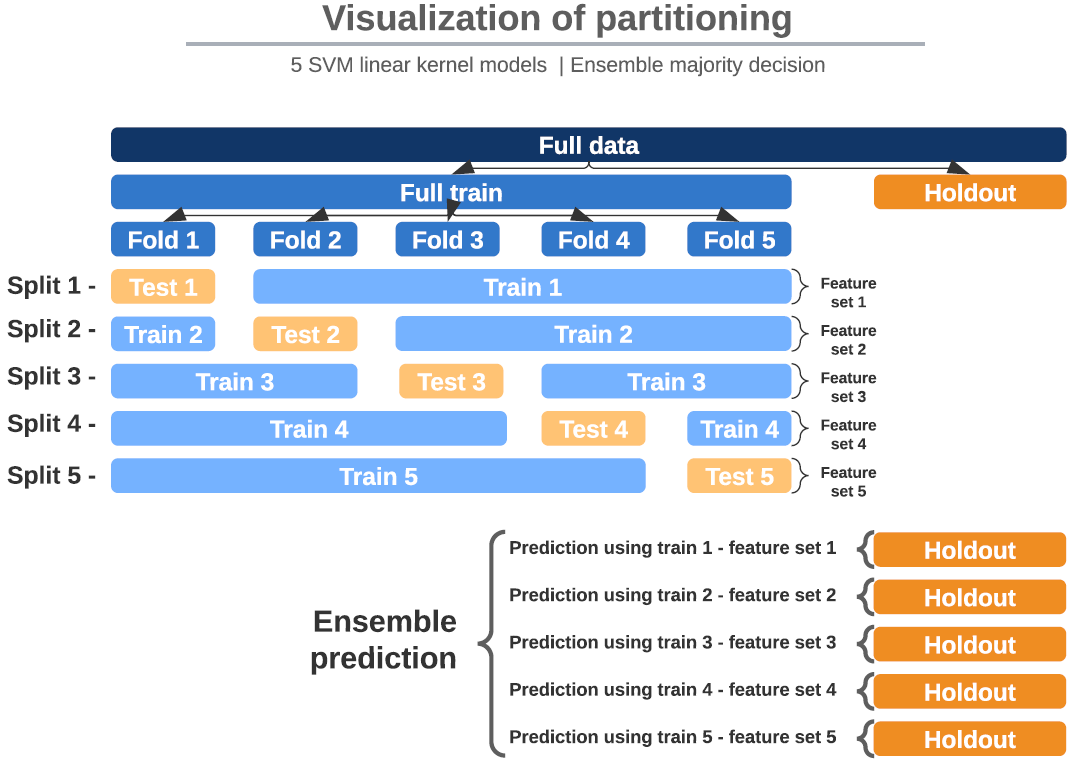


Figure 2.

*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. Moreover, performance was also calculated separately for the two sexes, to assess potential ML biases, given the different nature of voices between males and females.

## 2.2 Literature search for 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 Parola et al. in 2019 (Parola 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).

This search yielded an additional 709 papers that were manually screened for relevance by their title. Relevant papers – both from the meta-analysis and from the manual screening, were then explored by content, looking for papers that, 1) implemented ML to classify schizophrenia patients from healthy controls using acoustic features, 2) were transparent and well-documented, 3) were thorough in applying proper ML methods, 4) had larger amounts of data.

This narrowed the number of papers down to 8 papers (see appendix, 7.1). The study by Chakraborty et al. from 2018 was chosen for replication after carefully assessing relevant 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 participants went through 8 such trials that were recorded, except for in the study from 2014 by Bliksted et al., where the they also recorded 2 practice trials – meaning this dataset included voice recordings from 10 trials (Bliksted et al., 2014). This totaled in 1900 recordings with a duration mean of 18.18 seconds and a standard deviation of 14.84.   
\* kan man skrive det op sådan?? \*

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 ML algorithm proves to be versatile in its predictions across data sets. \* move to discussion? \*

### 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 |
| HC | 17 | 19 | 22.7 | 3.19 | 18-30 |
| Bliksted et al., 2014 | 46 | SZ | 6 | 17 | 23.3 | 3.94 | 18-33 |
| HC | 7 | 16 | 23.7 | 3.61 | 18-34 |
| Bliksted et al., 2019 | 48 | SZ | 11 | 8 | 40.8 | 12.4 | 20-61 |
| HC | 13 | 16 | 37.5 | 13.1 | 21-62 |
| Bliksted et al., n.d. | 58 | SZ | 12 | 18 | 24.8 | 3.66 | 18-31 |
| HC | 13 | 15 | 24.4 | 4.65 | 18-34 |
| Total | 106 | SZ | 45 | 61 | 26.7 | 9.02 | 18-61 |
| 116 | HC | 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. HC refers to the healthy control group, while SZ refers to the schizophrenic group.*

### 2.3.2 Task

The participants from all studies went through the Frith Happé animations task (Abell et al., 2000). 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. The videos differed their movement; some animations had the triangles move randomly, while others had either goal-directed movement or interactive movement between the triangles.

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

The toolkit openSMILE 2.3.0 was used for extracting the features needed for the 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 resulted 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. These feature vectors functioned as data points for the model.

## 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 \* introduction??? Or discussion? \* . 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. \*Discussion or introduction??\*

## 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 minimum and the maximum 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 (Myrianthous, 2020). This procedure is common practice when applying most ML algorithms. \* move to discussion or introduction? \*

## 2.7 Feature selection

### 2.7.1 L2 regularization

As the 988 acoustic features from the ‘emobase’ package were originally designed to distinguish emotions from speech, some features were less important 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 potentially superfluous features.

Feature selection was done using L2 regularization, also called the Least Absolute Shrinkage and Selection Operator (LASSO) analysis regression. LASSO regularization has the advantage of being able to shrink irrelevant parameters all the way to zero. To carry out this process, the ‘glmnet’ R Package was utilized for the purpose of this paper. (Friedman et al., 2010)

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 its value multiplied with a lambda value.

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:  
\* Brug eget lavet \*

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. \* which data was this from? \**

### 2.7.2 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, diagram

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.*

## 2.8 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, using only 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 previously mentioned constituent models. If for example 3 out of 5 models predicted ‘schizophrenia’ for a recording, then this was also the vote of the ensemble model.

## 2.9 Evaluation metrics

For evaluating the performance of the models, several metrics conveying information about the classification was provided. Information on precision, recall and F1-rates are provided for each class (controls and patients). A macro average F1-score is also provided for the models, as well as accuracy and baseline accuracy. Moreover, confusion matrices are 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.*

## 2.10 Differences between replication and original study

This replication employed principles from the proposed general pipeline, which means that it diverges from the original study on several aspects. These discrepancies can all be seen in table x \* below.

|  |  |  |
| --- | --- | --- |
|  | **Original** | **Replication** |
| **N (participants)** | 78 | 222 |
| **Origin** | Malay, Chinese, Indian | Danish |
| **SZ rate** | 66.67% | 48.2% |
| **Task  language** | English | Danish |
| **N (recordings)** | 78  (1 per participant) | 1900  (8-10 per participant) |
| **Mean (length of recordings)** | 26 min | 18.8 sec |
| **Feature  selection** | PCA | LASSO regularization |
| **Feature scaling** | Min-max  normalization | No  information provided |
| **ML  algorithm** | Single SVM | Majority  vote ensemble – SVM |
| **Final  testing set** | Cross- validation (full dataset) | Holdout (separate set for final test) |

Table x \* :  
*An overview of the differences between the original paper by Chakraborty et al. and this replication.*

# 3. Results

This section presents the performance of the ML 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 HC and SZ which mean ‘healthy controls’ 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 along with 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 |
| HC | 0.759 | 0.647 | 0.698 |
| Female | 70.00% | 51.11% | SZ | 0.689 | 0.705 | 0.697 |
| HC | 0.711 | 0.696 | 0.703 |
| Both | 70.32% | 51.87% | SZ | 0.675 | 0.739 | 0.706 |
| HC | 0.734 | 0.670 | 0.700 |

Table 3:  
*Performance of the ensemble model - within both the sexes and diagnosis.*

|  |  |  |  |
| --- | --- | --- | --- |
| N = 374  *(m = 194,*  *f = 180)* | Predicted group | | |
| True group |  | HC | SZ |
| HC | 130  *(m = 66*  *f = 64)* | 64  *(m = 36*  *f = 28)* |
| SZ | 47  *(m = 21*  *f = 26)* | 133  *(m = 71*  *f = 62)* |

Table 4:  
*Confusion matrix for the ensemble model predictions. Information on the proportion of males (m) and females (f) is also provided.*

# 4. Discussion

This discussion section will first compare the results of this replication with the results of the original paper. Focus will be on the metrics; precision, recall and F1-score. A potential model bias coming from the physiological difference between the sexes, will furthermore be investigated in relation to the results.

Secondly, the implementation of the general pipeline in this replication will be discussed – going into depth with the choices for each step. 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) with regards to this specific replication (evaluating the choices for the 9 steps) and  
2) with regards to the original paper (what differed in the replication, and what impact did it have?)

Finally, the use of the proposed general pipeline will be assessed, using the insights of this replication. Future research using the pipeline will also be discussed, looking into both the benefits and limitations that it holds, as well as potential future development. On this basis, the prospects for conducting further research are then assessed.

## 4.1 Performance results and comparison of performance to original study

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 had a baseline accuracy of 66.67% (2/3rd of the participants were patients), while this replication had a baseline accuracy 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 both the ensemble model and the model of the original study classified controls equally well. The model from the original study did, however, achieve a higher F1-score when classifying patients compared to the model from this replication. This performance difference is what caused the macro average F1-score to be higher for the study by Chakraborty et al. Moreover, both models also had an evenly balanced rate between recall and precision – the metrics that constitute the basis for the F1-score calculation.

As voice is modulated by the physiological differences between the sexes, it is relevant to see if this fact resulted in a model that predicted one sex better than the other. The ensemble model classified equally well between males and females with macro average F1-scores of 0.706 for males and 0.7 for females. No information was provided by Chakraborty et al. on this issue, although performance metrics would have been informative in shedding light upon a potential sex bias.

All performance measures considered, a moderate difference in performance was found with this replication seemingly having slightly worse classification capabilities. Should this fact be interpreted as an indicator of the original study not being conservative enough for the implementation of ML methods? Perhaps this rigorous ML pipeline implementation resulted in a more robust and generalizable model and its performance more accurately reflects what can be achieved. Although this could be hypothesized to partially explain the drop in performance, it is very likely that the differences across studies in data, feature scaling and feature selection also had an impact. To shed light on this topic it is sensible to evaluate the specifics of the individual pipeline steps in this replication, as well as discussing the differences between the study by Chakraborty et al. and this replication.

## 4.2 Evaluating specific pipeline implementation (and discussing differences from original study)

As the proposed general pipeline merely provides the overall workflow and not a rigid guide to the specific execution, the choices of each step have to be scrutinized and evaluated. Furthermore, the pipeline steps that deviate in nature from the original study, will also be discussed (for an overview of deviations, see table x \* ).  
**1) Data acquisition.** This study used data corpora of diverse speech recordings from multiple studies. Not only did this provide more data, but it also provided more diverse data since the recording setting differed across study. The ML model will therefore be more versatile, in that it is less bound to only learning patterns within a certain setting. \* måske slettes? \*

All participants in this study were Danish, which entails that the results are not necessarily entirely generalizable across nationalities or languages. As of yet, research suggests that symptoms such as alogia and the flat effect partially produce the differences in acoustic patterns in patients with schizophrenia. It is a probable that the extent to which they elicit these symptoms could be modulated by culture and language. Moreover, the nature of the schizophrenic participants might also vary slightly – people diagnosed with schizophrenia elicit different symptoms depending on culture, with for example westerners typically eliciting more depressive behavior (*Lundbeck Institute Campus*, 2016; Sartorius et al., 1986).

The number of recordings (N = 1900) was quite large in this replication given the large number of participants and the fact that each participant went through 8-10 trials with separate recordings. This meant that the feature extraction process produced a large number of feature vectors (1 per recording). As each feature vector represents a data point, the classification algorithm simply had many datapoints to learn from.  
The data acquisition was widely different in the original study. Their recordings were substantially longer which meant that the feature vectors for each data point more accurately captured the true feature values as they were less prone to random variation \* cite \*. However, they did only have 78 data points, substantially fewer than the 1900 employed in this replication. Their participants were Malay, Indian or Chinese, but were instructed to speak English during the recording sessions. As cognitive load has been known to show larger symptomatologic effects for voice in patients, this might have elicited stronger patterns for the model to pick up and correspondingly better predictions (Parola et al., 2019).  
As mentioned, both language and culture might impact both symptoms and acoustic measures in patients. This means that the results between studies can have been impacted by the difference in data acquisition.

**2) Preprocessing.** The data was cleaned, ensuring that for example reverb qualities or noises specific to certain rooms was not allowed to confound the classification algorithm. This sometimes proves to be a confounding factor in classification using speech, leading to high but inaccurate accuracies (Bone et al., 2013). The sound level of the data was normalized before and after the cleaning steps, to avoid having the ML model learn from the volume level. Reflecting upon this retrospectively, it does technically allow for the training data to learn from the holdout set, since this process happened before the splitting into a training and a holdout set. Given that loudness of speech is only one feature out many, it is expected that this have had a miniscule impact – if any.  
As for feature extraction, the ‘emobase’ feature set includes features originally found to be relevant for classifying emotions. Since emotional impairment is known to be one of the hallmark symptoms of schizophrenia, they also prove useful for classifying schizophrenia. However, many other acoustic features not included in ‘emobase’ have been found promising in this feat \*cite\*. It could also have been interesting to look at multiple different feature sets and compare them, or to use multiple in conjunction. The recent software development of ‘DigiVoice’ for example, supports feature extraction of not only acoustics features, but also features about linguistic complexity, natural language and semantic coherence features. (Zhang et al., 2019).

**3) Partitioning.** The data was partitioned into a training set and a holdout set of 80% and 20%, respectively. As mentioned in the introduction of the proposed general pipeline (section 1.3.2), the split ratio is somewhat arbitrary, although a larger training allows for the model to better learn the patterns in the data while a larger holdout set allows for a more accurate measure of performance.

It is, however, relevant to try and have both sets balanced. This study had a roughly even balancing of not only controls and patients, but also an almost even balancing between the sexes \* what balance – specifically? \*. The holdout set included enough male (N = 194) and female recordings (N = 180) to allow for insights into whether the slight imbalance in sex in the training data confounded the results. As discussed in 4.1.1 the model was inbiased in terms of sex.

**4) Feature scaling.** As using SVM as an algorithm requires scaled features, this study employed a min-max normalization. The scaling of both the training and holdout set used the minimum and maximum values from the training set to ensure no information could flow from the training to the holdout set (Myrianthous, 2020). An alternative to normalization would have been to standardize. Standardization has the benefit of not being as affected as outliers as min-max normalization does, given that standardized data is generated from standard deviation and mean.  
As no information was provided in the original paper, it is unclear whether their acoustic features were scaled within each step of the cross-validation, ensuring to scale the test set using only information from the training set, or if they scaled prior to the cross-validation process. The latter could result in a small amount of overfitting. Performance would appear to be slightly better, but it would more poorly reflect out-of-sample performance. The reason for this would be that the classification algorithm could have learned from the testing data before seeing it for the validation (Géron, 2019).

**5) Feature selection.** LASSO regularization was utilized for feature selection in this study. LASSO has in some studies found to be one of the best feature selection techniques, with great improvements of classification algorithms (Sun et al., 2019) \* cite more \*. However, a myriad of other feature selection techniques exists and could have been utilized instead. It would have been interesting to perform multiple feature selection techniques and compare performance as it was done in the original paper, instead of choosing only one technique.  
Principal Component Analysis (PCA) which was found superior in the original study has, similarly to LASSO, often been found to be the feature selection technique which resulted in the best ML performance when comparing models (Abdi & Williams, 2010; Chakraborty et al., 2018; Sun et al., 2019). Given that both PCA and LASSO have been known to perform similarly well, it is unlikely that all the variation in performance between the two studies can be attributed solely to feature selection technique. If the method for using the acoustic features from ‘emobase’ for classification truly is robust and reliable, then either should – at least in theory - work \* cite \*.

**6, 7, 8) Model training, model testing and parameter tuning.** SVM linear kernel models were utilized in the training step, as they were found to provide the best predictions in the original study. There are a range of alternatives, but using the exact same algorithm has the advantage of making comparison of models and results less challenging.   
The final majority vote ensemble model was, although, a bit different. Combining or utilizing multiple models within a single model has been seen to have benefits for performance and generalizability (Buracas & Albright, 1994; Hong & Page, 2004; Sechidis, 2020; Tang et al., 2006). The ensemble model can be hypothesized to give more robust results, but given it was only tested on a single holdout set these remain speculations and would require further testing across datasets.  
As a result of having both a test and a holdout set, hyperparameter tuning was possible - in contrast to the original study. Although the optimal hyperparameters values are not necessarily those found be the default software settings (Schratz et al., 2019), they were in this case, which meant that the tuning step could have been altogether skipped. As this conclusion was not available a priori, it was still a necessary step to test it out.

**9) Validation.** The performance was investigated primarily by the use of F1-scores and confusion matrices were provided for additional metrics to be calculated. Performance on the two sexes was also calculated to check for potential bias. Validation was quite similar in the original study, apart from the fact that they did not check for performance differences between subgroups. They employed different nationalities, and the fact that the participants came from different backgrounds might or might not have impacted the overall performance differently. Additional discussion of results is provided in section 4.1.  
Supplementary validation metrics could have been provided in the form of ROC-curves. These would have been able to show the tradeoff between false positives and false negatives at different classification thresholds. A change in threshold can be desirable depending on clinical application, and thus this information can be useful. Given that research has not yet established the robustness, generalizability and ecological validity of these ML algorithms, it was deemed to be too early to look at direct application. As a result, ROC-curves have been omitted.

As mentioned in the introduction, the implementation of the pipeline steps in solitude was not proposed to alleviate the issues in the current literature unaccompanied. The pipeline had to be accompanied by a proper and rigorous documentation. It also had to be supplemented by both reflection and scrutiny of the specific choices for each step in the pipeline. The description of the methods for this replication have been attempted to be both meticulous and exhaustive, enabling both replication and further scrutiny. The specific choices for each step have moreover been discussed here, both in terms of their consequences but also in terms of their potential alternatives.

In summary; the pipeline implementation of this replication is unlikely to have been confounded by problems related to either small or unbalanced data. This model is also unlikely to be overfit, given that the feature scaling process was carried out appropriately and given that the data was cleaned so that any acoustic qualities the model picked up on were not due to specific room settings. Moreover, the validation on the final holdout set was both balanced and potential bias coming from differences between the sexes were ruled out by looking at the results for each sex individually. The differences in performance between the original and this replication is likely to be due to a) a difference in data, b) a difference in methods, or c) the very conservative nature of this replication. Since the data differs widely between studies, it is reasonable to assume that data has had an impact – it is however unfeasible to deciphering exactly why, how and to which magnitude.

## 4.3 Limitations and prospects of the proposed overall pipeline

By providing a general pipeline for classification of schizophrenia patients, it is our hopes that the conditions for both replicability and comparisons of results can be improved. It is also the aim to alleviate future problems of overfitting within the literature. However, our proposal does not provide an exhaustive solution on its own. The pipeline does hold some limitations, some of which have become apparent through its use in this replication.

One of the limitations has to do with the broadness of the proposed pipeline. The pipeline was meant to broad and inclusive for it to be able to aid a variety of research endeavors. The generalist nature of the pipeline has a downside, however. Many of the choices for good ML conduct are still left up to the individual researcher which hosts room for error. Choices for algorithm, feature selection technique and model type are still left up to the practical experience of the individual researcher. The problem of knowing which algorithm to use for instance, is still apparent even given the use of the overall pipeline. The choice for choosing ensemble modeling and LASSO feature selection in this replication, for instance, was mostly based upon individual experience and knowledge of the existing methods. It is important to note, that although the proposed pipeline still does narrow down the number of potential choices to the most feasible choices.

Another limitation that has become apparent has to do with how difficult it can be to compare own results to other studies – even when applying thorough and conservative ML implementation. Using this replication as a case example, it is quite apparent that it proves difficult to pinpoint what drove the differences in performances. The drop in performance in this replication could be attributed to the difference in feature selection technique. It might also be attributed to the difference in data – participants came from different backgrounds, spoke different languages and were they were presented to task of dissimilar nature. The drop in performance could also be due to overfitting of the original study, since the documentation of their methods does not clearly specify whether feature scaling and feature selection was performed on the combined training and testing set or not. Testing across datasets could potentially be a solution - shedding more light on true out-of-sample performance.

A last self-evident limitation of the proposed solution has to do with how much the principles are applied. In order to pin down the usefulness of ML classification in schizophrenia, replications and original research using good ML conduct are all needed. This paper advocates for more conservative use of ML methods, proper documentation, a greater number of reproductions. Finally, this paper advocates for open-science, transparency of both methods and scripts with the intent of making research more cumulative.

# 5. References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459.

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

Andreasen, N. C., Arndt, S., Alliger, R., Miller, D., & Flaum, M. (1995). Symptoms of schizophrenia: Methods, meanings, and mechanisms. *Archives of General Psychiatry*, *52*(5), 341–351.

Bearden, C. E., Wu, K. N., Caplan, R., & Cannon, T. D. (2011). Thought disorder and communication deviance as predictors of outcome in youth at clinical high risk for psychosis. *Journal of the American Academy of Child & Adolescent Psychiatry*, *50*(7), 669–680.

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.

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of ‘Bias’ in NLP. *ArXiv:2005.14050 [Cs]*. http://arxiv.org/abs/2005.14050

Bone, D., Chaspari, T., Audhkhasi, K., Gibson, J., Tsiartas, A., Van Segbroeck, M., Li, M., Lee, S., & Narayanan, S. (2013). *Classifying Language-Related Developmental Disorders from Speech Cues: The Promise and the Potential Confounds*.

Buracas, G., & Albright, T. (1994). The Role of MT Neuron Receptive Field Surrounds in Computing Object Shape from Velocity Fields. In J. Cowan, G. Tesauro, & J. Alspector (Eds.), *Advances in Neural Information Processing Systems* (Vol. 6, pp. 969–976). Morgan-Kaufmann. https://proceedings.neurips.cc/paper/1993/file/ede7e2b6d13a41ddf9f4bdef84fdc737-Paper.pdf

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.

Claesen, M., & De Moor, B. (2015). Hyperparameter search in machine learning. *ArXiv Preprint ArXiv:1502.02127*.

Cohen, A. S., Mitchell, K. R., Docherty, N. M., & Horan, W. P. (2016). Vocal expression in schizophrenia: Less than meets the ear. *Journal of Abnormal Psychology*, *125*(2), 299–309. https://doi.org/10.1037/abn0000136

Cohen, A. S., Najolia, G. M., Kim, Y., & Dinzeo, T. J. (2012). On the boundaries of blunt affect/alogia across severe mental illness: Implications for Research Domain Criteria. *Schizophrenia Research*, *140*(1), 41–45. https://doi.org/10.1016/j.schres.2012.07.001

Corcoran, C. M., Mittal, V. A., Bearden, C. E., Gur, R. E., Hitczenko, K., Bilgrami, Z., Savic, A., Cecchi, G. A., & Wolff, P. (2020). Language as a biomarker for psychosis: A natural language processing approach. *Schizophrenia Research*.

Covington, M. A., He, C., Brown, C., Naçi, L., McClain, J. T., Fjordbak, B. S., Semple, J., & Brown, J. (2005). Schizophrenia and the structure of language: The linguist’s view. *Schizophrenia Research*, *77*(1), 85–98. https://doi.org/10.1016/j.schres.2005.01.016

DeVylder, J. E., Muchomba, F. M., Gill, K. E., Ben-David, S., Walder, D. J., Malaspina, D., & Corcoran, C. M. (2014). Symptom trajectories and psychosis onset in a clinical high-risk cohort: The relevance of subthreshold thought disorder. *Schizophrenia Research*, *159*(2), 278–283. https://doi.org/10.1016/j.schres.2014.08.008

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM Computing Surveys*, *27*(3), 326–327. https://doi.org/10.1145/212094.212114

Faber, J., & Fonseca, L. M. (2014). How sample size influences research outcomes. *Dental Press Journal of Orthodontics*, *19*(4), 27–29.

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

Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. In *Automated Machine Learning* (pp. 3–33). Springer, Cham.

Foody, G. M. (2017). Impacts of sample design for validation data on the accuracy of feedforward neural network classification. *Applied Sciences*, *7*(9), 888.

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

Géron, A. (2019). Feature scaling. In *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (pp. 69–70). O’Reilly Media. https://books.google.dk/books?hl=da&lr=&id=HHetDwAAQBAJ&oi=fnd&pg=PP1&dq=Hands-On+Machine+Learning+with+Scikit-Learn+and+TensorFlow&ots=0Lnl2wglVq&sig=ZdRI2rr1GjIiSpc764zQV-EMQDw&redir\_esc=y#v=onepage&q=As%20with%20all%20the%20transformations%2C%20it%20is%20important%20to%20fit%20the%20scalers%20to%20the%20training%20data%20only%2C%20not%20to%20the%20full%20dataset%20(including%20the%20test%20set).%20Only%20then%20can%20you%20use%20them%20to%20transform%20the%20training%20set%20and%20the%20test%20set%20(and%20new%20data)&f=false

Gosztolya, G., Bagi, A., Szalóki, S., Szendi, I., & Hoffmann, I. (2018). Identifying Schizophrenia Based on Temporal Parameters in Spontaneous Speech. *Interspeech 2018*, 3408–3412. https://doi.org/10.21437/Interspeech.2018-1079

Guzzetta, G., Jurman, G., & Furlanello, C. (2010). A machine learning pipeline for quantitative phenotype prediction from genotype data. *BMC Bioinformatics*, *11*(8), S3. https://doi.org/10.1186/1471-2105-11-S8-S3

Hitczenko, K., Mittal, V. A., & Goldrick, M. (2020). Understanding Language Abnormalities and Associated Clinical Markers in Psychosis: The Promise of Computational Methods. *Schizophrenia Bulletin*. https://doi.org/10.1093/schbul/sbaa141

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

Hutter, F., Hoos, H., & Leyton-Brown, K. (2014). An Efficient Approach for Assessing Hyperparameter Importance. *International Conference on Machine Learning*, 754–762. http://proceedings.mlr.press/v32/hutter14.html

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

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Jović, A., Brkić, K., & Bogunović, N. (2015). A review of feature selection methods with applications. *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1200–1205.

Kliper, R., Portuguese, S., & Weinshall, D. (2016). Prosodic Analysis of Speech and the Underlying Mental State. In S. Serino, A. Matic, D. Giakoumis, G. Lopez, & P. Cipresso (Eds.), *Pervasive Computing Paradigms for Mental Health* (pp. 52–62). Springer International Publishing. https://doi.org/10.1007/978-3-319-32270-4\_6

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, *117*(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering*, *160*(1), 3–24.

Kuperberg, G. R. (2010). Language in Schizophrenia Part 1: An Introduction. *Language and Linguistics Compass*, *4*(8), 576–589. https://doi.org/10.1111/j.1749-818X.2010.00216.x

*Lundbeck Institute Campus*. (2016, January 6). https://institute.progress.im/en/content/schizophrenia-across-cultures

Mantovani, R. G., Horváth, T., Cerri, R., Vanschoren, J., & de Carvalho, A. C. (2016). Hyper-parameter tuning of a decision tree induction algorithm. *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, 37–42.

Martínez-Sánchez, F., Muela-Martínez, J., Cortés-Soto, P., Meilán, J., Ferrándiz, J., Egea-Caparrós, D., & Valverde, I. M. (2015). Can the Acoustic Analysis of Expressive Prosody Discriminate Schizophrenia? *The Spanish Journal of Psychology*, *18*, 1–9. https://doi.org/10.1017/sjp.2015.85

Morice, R. D., & Ingram, J. C. (1983). Language complexity and age of onset of schizophrenia. *Psychiatry Research*, *9*(3), 233–242.

Myrianthous, G. (2020, June 28). *Feature Normalisation and Scaling | Analytics Vidhya*. https://medium.com/analytics-vidhya/feature-scaling-and-normalisation-in-a-nutshell-5319af86f89b

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

Olson, R. S., La Cava, W., Mustahsan, Z., Varik, A., & Moore, J. H. (2017). Data-driven advice for applying machine learning to bioinformatics problems. *ArXiv Preprint ArXiv:1708.05070*.

Olson, R. S., & Moore, J. H. (2016). TPOT: A tree-based pipeline optimization tool for automating machine learning. *Workshop on Automatic Machine Learning*, 66–74.

Oreski, D., Oreski, S., & Klicek, B. (2017). Effects of dataset characteristics on the performance of feature selection techniques. *Applied Soft Computing*, *52*, 109–119. https://doi.org/10.1016/j.asoc.2016.12.023

Parola, A., Simonsen, A., Bliksted, V., & Fusaroli, R. (2019). *Voice Patterns in Schizophrenia: A systematic Review and Bayesian Meta-Analysis* [Preprint]. Bioinformatics. https://doi.org/10.1101/583815

Püschel, J., Stassen, H. H., Bomben, G., Scharfetter, Ch., & Hell, D. (1998). Speaking behavior and speech sound characteristics in acute schizophrenia. *Journal of Psychiatric Research*, *32*(2), 89–97. https://doi.org/10.1016/S0022-3956(98)00046-6

Rapcan, V., D’Arcy, S., Yeap, S., Afzal, N., Thakore, J., & Reilly, R. B. (2010). Acoustic and temporal analysis of speech: A potential biomarker for schizophrenia. *Medical Engineering & Physics*, *32*(9), 1074–1079. https://doi.org/10.1016/j.medengphy.2010.07.013

Samad, M. D., & Witherow, M. A. (2018). A Machine Learning Pipeline to Optimally Utilize Limited Samples in Predictive Modeling. *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–5. https://doi.org/10.1109/ICCCNT.2018.8494100

Sanders, S., & Giraud-Carrier, C. (2017). Informing the use of hyperparameter optimization through metalearning. *2017 IEEE International Conference on Data Mining (ICDM)*, 1051–1056.

Sartorius, N., Jablensky, A., Korten, A., Ernberg, G., Anker, M., Cooper, J. E., & Day, R. (1986). Early manifestations and first-contact incidence of schizophrenia in different cultures: A preliminary report on the initial evaluation phase of the WHO Collaborative Study on Determinants of Outcome of Severe Mental Disorders. *Psychological Medicine*, *16*(4), 909–928. https://doi.org/10.1017/S0033291700011910

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, *406*, 109–120. https://doi.org/10.1016/j.ecolmodel.2019.06.002

Sechidis, K. (2020, March 5). *A Machine Learning perspective on the emotional content of Parkinsonian speech*. https://www.youtube.com/watch?v=X-8V8FT-M78&feature=emb\_title&ab\_channel=AppliedMachineLearningDays

Sichlinger, L., Cibelli, E., Goldrick, M., & Mittal, V. A. (2019). Clinical correlates of aberrant conversational turn-taking in youth at clinical high-risk for psychosis. *Schizophrenia Research*, *204*, 419–420. https://doi.org/10.1016/j.schres.2018.08.009

Solomon, M., Olsen, E., Niendam, T., Ragland, J. D., Yoon, J., Minzenberg, M., & Carter, C. S. (2011). From lumping to splitting and back again: Atypical social and language development in individuals with clinical-high-risk for psychosis, first episode schizophrenia, and autism spectrum disorders. *Schizophrenia Research*, *131*(1–3), 146–151.

Stassen, H. H., Albers, M., Püschel, J., Scharfetter, Ch., Tewesmeier, M., & Woggon, B. (1995). Speaking behavior and voice sound characteristics associated with negative schizophrenia. *Journal of Psychiatric Research*, *29*(4), 277–296. https://doi.org/10.1016/0022-3956(95)00004-O

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

Sun, P., Wang, D., Mok, V. C., & Shi, L. (2019). Comparison of feature selection methods and machine learning classifiers for Radiomics analysis in glioma grading. *IEEE Access*, *7*, 102010–102020.

Tahir, Y., Yang, Z., Chakraborty, D., Thalmann, N., Thalmann, D., Maniam, Y., binte Abdul Rashid, N. A., Tan, B.-L., Lee Chee Keong, J., & Dauwels, J. (2019). Non-verbal speech cues as objective measures for negative symptoms in patients with schizophrenia. *PLoS ONE*, *14*(4). https://doi.org/10.1371/journal.pone.0214314

Tang, E. K., Suganthan, P. N., & Yao, X. (2006). An analysis of diversity measures. *Machine Learning*, *65*(1), 247–271.

Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013). Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 847–855.

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLOS ONE*, *14*(11), e0224365. https://doi.org/10.1371/journal.pone.0224365

Voleti, R., Woolridge, S., Liss, J. M., Milanovic, M., Bowie, C. R., & Berisha, V. (2019). Objective Assessment of Social Skills Using Automated Language Analysis for Identification of Schizophrenia and Bipolar Disorder. *ArXiv:1904.10622 [Cs]*. http://arxiv.org/abs/1904.10622

Zhang, L., Chen, X., Vakil, A., Byott, A., & Ghomi, R. H. (2019). DigiVoice: Voice Biomarker Featurization and Analysis Pipeline. *ArXiv:1906.07222 [Cs, Eess]*. http://arxiv.org/abs/1906.07222

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

# 6. Appendix

## 6.1 Relevant studies

(Chakraborty et al., 2018; Gosztolya et al., 2018; Kliper et al., 2016; Martínez-Sánchez et al., 2015; Püschel et al., 1998; Rapcan et al., 2010; Stassen et al., 1995; Tahir et al., 2019)

## 6.2 Feature lists after L2 Regularization

\*