To do list:

1. Tæller acknowledgements med I character count? If no add: Riccardo Fusaroli for supervising this project + Bliksted et al. for providing the data.
2. Tæller formler som 800 tegn? Eller Ikke?
3. Spørg Riccardo om det fra StackExchange. Også om det er okay uden konklusion? Også om det overhovedet kan kaldes replication? Og om formler tæller 800 tegn? Eller de tegn de er af. Og om acknowledgements tæller med
4. Tjek eventuelt svar fra Fabio
5. Spørg Anne og Aske om de vil læse igennem?
6. Send til Alberto Parola. (Husk at CC Riccardo)
7. Send til Tour de Bachelor (i gruppechatten)
8. Send til Stinne/Mads (som PDF)
9. RICCARDOS NOTES:
10. “*- make sure to have somebody else read the paper and point out where things are "indforstået", and generally give an eye on how fluent the language is. (hard to focus on that when you're still working on results and structure). Nothing major, but it does make a difference w censors.*”
11. *“- feel free to send a more complete draft to Alberto Parola <a.parola@unito.it> for comments (cc'ing me and mentioning I said to do that). More complete in that he's not been part of the process, so he'll need a bit more to be able to follow.*”

Feature selection deleted:  
*The latter is a penalty based on the size of the beta estimate multiplied by a lambda value.*

*The loss function used for finding parameter estimates using LASSO:  
*

*To find the optimal lambda value, lambda.min value was first computed by testing a range of values using 5-fold cross-validation and finding the value that minimizes the LASSO loss function. Subsequently, the lambda value resulting in the fewest number of features within 1 SE from lambda.min was chosen, as it acknowledges that fits are estimated with some error (Friedman et al., 2010).*

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 a range of 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.*



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. The plot is from the first training set.*

This process thus generated a list of only the relevant features for classification of patients from controls. For a visualization of lambda misclassification plot, see figure 3.

# Abstract

This study replicated two promising ML studies on new data, using an improved validation technique and an inclusion of sensitivity and specificity rates. Accuracy rates found through replication were dissimilar to the original studies, with study X\* and study Y\* having overall accuracy rates for classification at 60% and 67%. In other words, a drop of 6 and 3 percentage points for the two studies, respectively. Through discussion, this study has found that the difference in scores in the replication points toward low ecological validity and robustness. The rest of the literature was also discussed, and I found that the widely heterogeneous results within the field indicate similar trends.  
As a consequence, this study has attempted to establish a ML pipeline less prone to the pitfalls of ML, with the intention of establishing a general procedure for future research. Finally, this paper advocates for a more open and cumulative scientific community.

How to write abstract:  
<https://blackboard.au.dk/bbcswebdav/pid-2793891-dt-content-rid-9152972_1/courses/BB-Cou-Hold-36086/L1%20-%20Getting%20started.pdf>  
p. 14 - p. 18

# 1. Introduction

## 1.1 Schizophrenia and biomarkers

### 1.1.1 Schizophrenia

**Actual paper:**

**Notes for section:**

Schizophrenia is a diagnosis which has long been defined by disturbances in both thought, perception and communication (Bleuler, 1950). Although schizophrenia has been known to be a group of great heterogeneity, patients oftentimes suffer from similar symptoms (Picardi et al., 2012; Tsuang et al., 1990). They are generally thought to be divided up into two types of symptoms; negative and positive. Positive symptoms are those that are present during a psychotic episode in schizophrenia and include delusions and hallucinations (Andreasen et al., 1995).   
  
Negative symptoms are those that either diminishes or halts thought processes or normal emotional functioning and include, but are not limited to asociality, alogia – poverty of speech, latency of speech and blocking, and blunted affect – a decrease in emotional expression and a lack of vocal intonation (Andreasen et al., 1995; Cohen et al., 2012).

Schizophrenia is furthermore associated with several other speech impairments in addition or in relation to the qualitatively described symptoms of alogia and blunted affect.

. Symptoms are qualitatively described with terms such as alogia and blunted affect – referring to characteristics such as poverty of speech, latency of speech, blocking, a decrease in emotional expression and a lack of vocal intonation (Andreasen et al., 1995; Cohen et al., 2012). 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 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.

### 1.1.2 Biomarkers and voice atypicalities

**Actual paper:**  
  
**Notes for section:**

Biomarkers - why voice good? -> shows part of social impairment.

1. Schizophrenia and voice in general
   1. 3 methods of studying
2. Two meta-studies -> the lit. is a mess

1.

History of the project:

Schizophrenia has certain distinctive features vocally. (Alogia, blunt affect, "poverty of speech", "latency of speech" etc.). This has been known since forever (Bleuler, 1911; Kraepelin, 1919).

 Voice atypicalities in SZ’s have always been known (Bleuler, 1911; Kraepelin, 1919).  
Schizophrenia has certain distinctive features vocally. Qualitatively the atypicalities have been described using numerous different terms (Alogia, blunt affect, "poverty of speech", "latency of speech", increased pauses, distinctive tone, intensity of voice etc.).

Voice atypicalities have been studied using 3 methods. Qualitative perceptual ratings, quantitative acoustic analysis and ML investigations.

Qualitative perceptual ratings have found robust differences between SZ and TD.

Quantitative acoustic analyses have found fewer robust differences, with varying effect sizes and sometimes direction.

Quantitative acoustic analyses have identified acoustic features on the basis of automated processes, leaving the assessment of the acoustic features more reliable. Using automation, the features of a set of voice data will identical over multiple feature detections, given the same feature detection hard- and software.  
Here, fewer robust differences were found with varying effect sizes and direction, depending on the features investigated (Cohen et al., 2014; https://www.biorxiv.org/content/10.1101/583815v4.full.pdf).

This is also in next section?:

Multivariate ML investigations have found promising results. Focus on minimizing out-of-sample-error instead of within sample-error as when using more traditional analyses, makes the applicability of the method more practically generalizable. It also allows for analyzing multiple features in conjunction. High correlation between almost all features (3.3, correlation <https://www.biorxiv.org/content/10.1101/583815v4.full.pdf> ).  
It does, however, not allow for transparency as to wherein the acoustic differences between SZ and HC lie.

2.

We don't know which features proves to have differences between SZ and TD

The litt. is a mess - results in different directions.

There's already a metastudy on Schizophrenia; which found atypicalities on different voice/speaking parameters - with varying effect sizes.

Large heterogeneity between studies.

More demanding tasks meant larger effect sizes.

## 1.2 Machine learning for detection of acoustic patterns

### 1.2.1 Prospects of machine learning in classifying schizophrenia

**Actual paper:**  
  
**Notes for section:**

1. Metatext and motivation for going into depth with machine learning
2. Allows for finding features (feature selection)
3. Allows for analyzing multiple features in conjunction
4. Promising findings (high accuracy in many studies)
5. Less interpretability but more practical applications (cheap)

Will not go into other ML things (gesticulation or others – beyond the scope of this paper)

1.

Although:  
Qualitative perceptual ratings have found relatively robust differences in voice between SZ and TD. Relying on raters to assess perceptual differences has some limitations. A feature such as “latency of speech” is interpretable and is partly going to be rated on the basis of human intuition – this requires comprehensive training for the rater. Moreover, the complex interplay between multiple acoustic features is hardly very accessible, even given proper and rigorous training.  
  
Therefore:

ML doesn’t have this problem.

2.

Feature selection; ridge, lasso, elasticnet

3.

Multivariate ML investigations have found promising results. Focus on minimizing out-of-sample-error instead of within sample-error as when using more traditional analyses, makes the applicability of the method more practically generalizable. It also allows for analyzing multiple features in conjunction. High correlation between almost all features (3.3, correlation <https://www.biorxiv.org/content/10.1101/583815v4.full.pdf> ).

4.

Multivariate ML investigations have found promising results. Focus on minimizing out-of-sample-error instead of within sample-error as when using more traditional analyses, makes the applicability of the method more practically generalizable. It also allows for analyzing multiple features in conjunction. High correlation between almost all features (3.3, correlation <https://www.biorxiv.org/content/10.1101/583815v4.full.pdf> ).

5.

Clinical application -> given schizophrenia, and given samtaleterapi or drugs, see how they're doing along the way by them talking every week on their phone.  
"*In addition, voice analysis may potentially allow to assess the response to psychosocial or pharmacological treatment over longer periods using objective and quantitative indices, and enhance the capability of clinicians to capture the complex relationship between emotion regulation, expressive behavior, social perception and cognitive and clinical features of the disorder (e.g. Ben-Zeev et al., 2017; Dahlgren et al., 2018; Tahir et al., 2019)*" (Parola, Fusaroli et. al 2019)Va [(Bush et al., 1998)](https://www.zotero.org/google-docs/?KFKj12).  
  
It does, however, not allow for transparency as to wherein the acoustic differences between SZ and HC lie.

ML can perhaps help with showing:  
a) Severity of schizophrenic symptoms  
b) Diagnosis, schizophrenia

Practical applications:   
Assisting tool for assessing diagnosis (Parola, Fusaroli et. al 2019)

On prediction of severity of clinical features from acoustic measures:  
(Püschel et al., 1998)

6 (extra):

Applicability of Bachelors project:

Meta-science, open science.

Assisting tool for assessing diagnosis (Parola, Fusaroli et. al 2019)

Clinical application -> given schizophrenia, and given samtaleterapi or drugs, see how they're doing along the way by them talking every week on their phone.

"*In addition, voice analysis may potentially allow to assess the response to psychosocial or pharmacological treatment over longer periods using objective and quantitative indices, and enhance the capability of clinicians to capture the complex relationship between emotion regulation, expressive behavior, social perception and cognitive and clinical features of the disorder (e.g. Ben-Zeev et al., 2017; Dahlgren et al., 2018; Tahir et al., 2019)*" (Parola, Fusaroli et. al 2019)

Companies interested in this (Lasse Hansen), Switzerland Internship on this in depression

### 1.2.2 Current limitations in the literature

**Actual:**

**Notes:**

* Overfitting
* Differences in methods, method quality and levels of transparency
  + Effect sizes of acoustic features is partially determined by task (difficulty)
* Lack of replications
  + Promising results
  + No validation between datasets
  + Language differences
  + No performance robustness measures

## 1.3 Alleviating current limitations

### 1.3.1 Through replications

**Actual paper:**  
  
**Notes for section:**

s

### 1.3.2 Through proper ML implementation

**Actual paper:**  
  
**Notes for section:**

*PCA reduces the dimensionality (number of features) of each data point (each recording), by generating a smaller number of new ‘principal components’ (dimensions) while preserving as much as the variation in the data as possible (Abdi & Williams, 2010). The latter feature selection technique diminishes the interpretability of the model as opposed to the former, given that the original acoustic features are convoluted in the new principal components. LASSO allows for investigations into which features where most important for classification.*

1. Meta text to have a rød tråd
2. Establish the need for a change in research (from above section)
3. Idea of a pipeline
4. Explain how a pipeline might alleviate the problem. And cause the change that is needed

Set up a standard pipeline.  
(Very) general introduction to pipeline in introduction. Just the conceptual structure.  
  
In intro:

“explain the most basic and most important steps”

BUT … (not feature selection on full dataset) etc.

Mention confusion matrix

In methods:  
Take the individual conceptual steps from the intro, and explain in specific detail how we tailored this ML “to the conceptual steps”.

The proposed pipeline follows a relatively simple overall structure, which can be thought of as 8 steps. It is important to note, that the course of action within each step to a certain extent depends upon the specific data and classification problem. Moreover, it is important to note that documenting extensively throughout is important for increasing transparency. High transparency a) allows for future replications and b) increases the applicability of the work, since knowledge will be available about exactly under which conditions a given machine learning approach performs a certain way. The steps of the pipeline are as follows:   
**1) Data acquisition.** Acquiring the voice data for fitting a machine learning algorithm – this may be data that has been acquired either through recording, or from using already recorded data. **2) Data preprocessing.** Consists of cleaning the data; cutting away irrelevant speech (e.g. from interlocutors), noise removal and extracting features. It might also include applying data augmentation (e.g. applying noise, reverb etc.). **3) Data partitioning.** Partitioning the data into a training and a holdout set for testing. Typically, a split of roughly 80/20 is used. Generally, a larger holdout set means more precise knowledge of the performance of a given model. A larger training set on the other hand generally means better and more robust predictions. **4) Feature scaling and selection.** Feature scaling is necessary for all machine learning algorithms that are distance-based. In these cases, scale the train and holdout set separately. Use information from the training set (e.g. min. and max. values in min./max. normalization) to do the scaling on both sets. This ensures no leakage of information from the training set to the test set, while still providing a common scale without losing information or distorting the differences in the range of values. Feature selection entails selecting only the relevant machine learning parameters, that improves classification. Many different methods can be applied in order to obtain a relevant feature set. It may also be skipped if theory or some motivation dictates it. However, feature selection tends to generally improve the robustness of predictions, shorten training times, avoid the curse of dimensionality and it increases interpretability. **5), 6), 7) Model training, parameter tuning, model testing.** These three steps are intertwined, and the cycle may be repeated. Divide the training set up into training and test and then train and test the model. After seeing how it performs on the test set, you may tune the parameters and repeat the process until the predictions are optimal. This can be done advantageously by the implementation of cross validation. **8) Validation on hold-out set.** The model then predicts the holdout set and performance is evaluated using relevant metrics. To allow for further insights into the performance of a given model, evaluation metrics ought also to be calculated for subgroups of participants (e.g. men/women or nationalities).

An overview of the pipeline can be seen in figure 1.

Fig 1.

*Pipeline with workflow that is well suited for classification machine learning on voice.*

**Notes for section:**

* S

s

### 1.3.3 Thesis statement / purpose of paper

Short summary of introduction

1. Voice is an important biomarker with practical applications if automated
2. Machine learning proves promising but there are issues with:
   1. Overfitting
   2. Difference in implementation making it impossible to specify what works
   3. Lack of replications and testing across datasets
3. Pipeline alleviates problems of
   1. 1) overfitting
   2. 2) difference in implementation
   3. 3) lack of replications (by making it easier)

Thesis statement:

1. Provide pipeline
2. Show example of implementation
3. Evaluate implementation

**Actual paper:**  
  
**Notes for section:**

Noone should just mindlessly replicate, when replicating!! Important to take all the necessary steps.

Overall goal of thesis:  
How does it work to reproduce?  
Increase conservative  
What are the important lessons we’re learning while we’re replicating

**Riccardos words on overall goal of thesis:**

How does it work to replicate the results of these paper when you increase the conservativeness of the procedure - and what do you learn from the problems, that future researcher can use when they do these kind of analysis

**Thesis statement idea 1 (Maries):**

This thesis aims to investigate the capabilities of existing machine-learning classifying individuals with ASD from acoustic features. We will review previous literature, extract strong voice-features and machine-learning models, and validate models on new data. We predict that support vector machine will achieve higher accuracy but will have less x and that naive bayes will x. Additionally, we predict that validation methods x,y,z will make results stronger in specific case/weaker generalization. By this, we will attempt to establish a procedure for machine-learning studies that achieve the most robust and ecologically valid measures.

**Thesis statement idea 2:**

This thesis aims to replicate two promising findings of machine learning classification of schizophrenia, using voice data. Since the literature on the area has very heterogeneous findings, I expect worse performance given the new data that I will test on. Given the inrobustness and low ecological validity of ML attempts, I will attempt to establish a ML pipeline less prone to the pitfalls of ML, with the intention of establishing a general procedure for future research.

Tour de Bachelors:

[https://docs.google.com/document/d/1qc3tDtAg6sc2-zfnxaxqKl\_WydK3AAgaDLb1V7QjTTU/edit?fbclid=IwAR1JB53UmJcDEI8GnXEEvA4PcuWXvVeX\_ZN43VEamHHxMWsHYdAR\_Wo3vKY#](https://docs.google.com/document/d/1qc3tDtAg6sc2-zfnxaxqKl_WydK3AAgaDLb1V7QjTTU/edit?fbclid=IwAR1JB53UmJcDEI8GnXEEvA4PcuWXvVeX_ZN43VEamHHxMWsHYdAR_Wo3vKY)

Pipeline:

[https://docs.google.com/document/d/1fbfpR5ZQiVTZYChzWut06fkXA9CQziMMJNtFOiF6Giw/edit?fbclid=IwAR3MU3OTehQ\_nVEk0nB8PihR\_0clhxrIbOPE5y\_v6X9PQoqRfFKRbZDiF7o#](https://docs.google.com/document/d/1fbfpR5ZQiVTZYChzWut06fkXA9CQziMMJNtFOiF6Giw/edit?fbclid=IwAR3MU3OTehQ_nVEk0nB8PihR_0clhxrIbOPE5y_v6X9PQoqRfFKRbZDiF7o)

# 4. Discussion

## 4.1 Results and replication comparison

### 4.1.1 Performance

**Actual:**

**Notes:**

### 4.1.1 Performance comparison to original study

1. Performance of models on test
   1. Evaluation metrics (F1, accuracy, precision + recall,)
      1. F1-score for model overall
      2. F1-scores for patients and controls respectively
      3. Accuracy (not very telling)
      4. Precision + Recall
   2. Between sexes
      1. Well balanced in replication
      2. No information in original paper
         1. Ought to be included
   3. Wrap up about results
      1. Similar results, but slightly better predictions in original
2. Why are there performance differences?
   1. Methods (as will be discussed in next section)

*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.*

Maybe include:

1. Performance of models on train
   1. High performance
   2. Low generalizability due to overfitting
      1. Mention bad study that overfits

* These things in terms of what is being predicted (train/test/holdout) - better predictions when overfitting – obviously. To underline how much more shit the predictions will get when you’re rigorous. With increased conservatism, how good are the results really?
* Performance between groups
  + Sexes:
    - This: Balanced test-set (but perhaps a bit worse training)
    - Original: No information. But balanced dataset.
  + Nationalities
    - This: None
    - Original: No information. But unbalanced dataset
* Where do the differences come from?
  + Data, Feature selection, Methods
  + See next sections

#### 4.1.2.1 Data

* Language/nationality
  + Biased because of difference in labeling
    - This: Danish diagnostics
    - Original: Chinese, Malay, Indian diagnostics
  + Biased because of difference in language
    - This: Danish
    - Original: 3 Countries, with different languages
* Task
  + This: mid-level difficulty; description of triangles. No social component
  + Original: high-level difficulty; interview. Social component
* Data quantity
  + This: More participants with shorter recordings
  + Original: Fewer participants with longer recordings
* Sound quality
  + This: Difference in recording equipment
  + Original: Maybe?
* What contributed to the differences in performance? (If any)
  + Possibly all. Likely not sound quality to a large extent
* The difference in the participants native country meant that not only did the language spoken in the recordings differs, but also that they were not speaking their own native language. Moreover, the pool of schizophrenic participants was likely to vary between the original and this replication. This is because both diagnostic tools and psychologist and psychiatrist training are heterogeneous between countries to some extent. \* CITE \*. \*PASSER DET??\*

#### 4.1.2.2 Partitioning

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

#### 4.1.2.3 Feature scaling

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

#### 4.1.2.4 Feature selection

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

Explanation of how it could be understood:

“*the features of the training set were ranked using one of the following techniques: F-score (ANOVA), χ 2 , Mutual Information, Pearson correlation, Principal Components, linear SVM, Decision Trees, and Random Forests. Subsequently, the optimal number of features were selected according to the previous ranking methods*”  
PCA used to rank? Most common method is that PCA is used for defining new features, namely PC1 + PC2 + ... +PCn, until some desired threshold of accumulated variance is met.

There’s also the possibility that it truly was used to rank, e.g. by looking at the features with least shared variance in the different principal components to avoid covarying features, but also here it is not possible to replicate 1-1. The method is still not specified

Shouldn’t really matter:

Regardless, of method used by Chakraborty et al, the method used here is good. And if the method using speech for classification truly is robust, then either would work. If these results truly are reliable and reliable, they shouldn’t be dependent on PCA/LASSO / whatever

Understanding PCA notes:

Link of idea of PCA for feature selection. (starts at 3:50). It shows that there are different methods (example with gain, here)

<https://www.youtube.com/watch?v=YEDOSOd44bU&list=PLBv09BD7ez_5_yapAg86Od6JeeypkS4YM&index=2&frags=wn&ab_channel=VictorLavrenko>

Link for example of PCA for feature selection (creating new features):

<https://www.quora.com/How-do-you-use-PCA-for-feature-selection>

#### 4.1.2.5 Machine learning algorithm

* Predicting (single participants, or same participants multiple times)
  + This: Predicting .wav files (several for each participant)
  + Original: Predicting participants
  + Should not have large impact on performance
* Ensemble modeling vs. Single machine learning algorithm
  + Stacking ensemble modeling
    - Better (if models are diverse, and generally good)
    - Only very slightly better
  + Single machine learning algorithm
    - Slightly worse

Should not have large impact on performance

Creating an ensemble model as opposed to using a single algorithm has the advantage of (possibly \* ) being more robust and reliable in its out-of-sample predictions (as mentioned in 4.1.1). Similarly, having the final testing set being

### 4.1.4 Methods ()

**Actual:**

**Notes:**

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

MAYBE INCLUDE:

* Specifics on ensemble modeling
  + Diversity/data trade-off in ensemble modeling

FOR BELOW I DO THE OPPOSITE NOW!!!!

Diversity/data trade-off in ensemble modeling:

I use all training data in each of the ensemble-sub-models. As opposed to excluding the test sets, that were also excluded for feature selection.

Could this be an issue?  
Yes; groups of diverse problem solvers (in general) outperform, the best (also often similar) models. At least when the diverse problem solvers and the better, more similar models have roughly the same amount of data. (Hong & Page, 2004)

Why did I choose to do it anyways?  
The “diversity” the opposing idea would bring, is not due to difference in neither type of models or any other diversity parameter. The opposing choice would only give diversity from differences in training data.

The increase in diversity would in this study, be on the cost of less training data. And less training data means worse predictions in general.

Does it really matter?

The trade-off between more/less training data and more/less diversity is unlikely to have had much of an impact. E.g. Less than 2 percent increase in acc. when having 10 agents (and we only have 5, which would probably mean even less of an impact) (Hong & Page, 2004).

But the potentially very small positive effect a more diverse set of decision-agents, might very well be negated by the fact that all of the 5 diverse models would be worse, due to their more limited data. In other words; no – it isn’t likely to have had a large effect. But it would have been interesting to do both.

## 4.2 Pipeline

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

**Actual:**

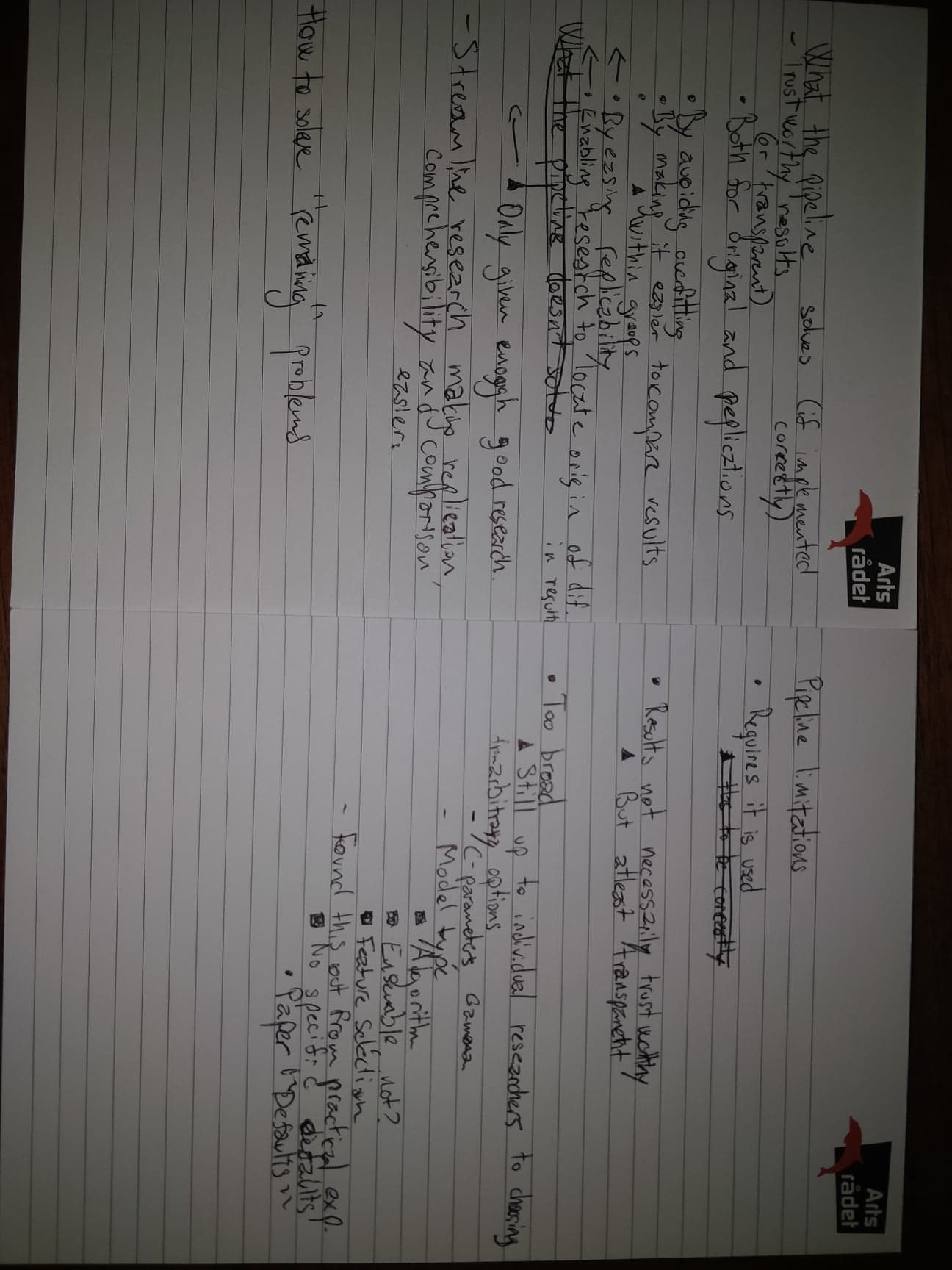
**Notes:**

*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.*

1. What the pipeline solves (if it is followed)
   1. Gives trustworthy (or at least transparent) results
      1. Avoiding overfitting
      2. Rigorous documentation
   2. Easier to compare results
      1. Conf. matrices
         1. Getting more information (e.g. on recall or others)
      2. Within groups
   3. Streamlining research
      1. Making replications easier
      2. Better comprehensibility
      3. Comparisons easier
   4. Potentially enabling research to locate origins of differences in results, as everything is documented
      1. Easier to use exact same pipeline, and only changing 1 thing (e.g. different language)
2. Pipeline limitations

The data was cleaned to prevent reverb qualities or noises specific to certain rooms to confound the study. 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.

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



1. Pipeline when used for replication
   1. If difference in results, hard to pinpoint
   2. Only possible given proper documentation
   3. Comparison only possible given proper documentation
3. Replication
   1. Possible but feature selection not so much
      1. (Methods explained in condensed manner in original)
4. Comparison of evaluation
   1. Good, but somewhat deficient
      1. More information on sexes and nationalities needed
5. Replication got similar results
   1. Slightly different
   2. Slight difference in performance – where from?
      1. Biased labels
      2. Difference in language
      3. Task differences
      4. Difference in algorithms
      5. Arbitrary choices for tuning
      6. A mixture (which mixture?) of all the above
   3. Some things might balance each other’s out, some might not
6. Reflection + proper documentation
   1. Yes – remembered to do this
7. Wrap up – Replication seems to have worked out OK

## 4.3 Further research

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

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

### 4.3.2 Benefits and limitations of the use of this pipeline in further research and going forward (Wrap-up)

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

A generally more open-science based approach

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

Differences to original study? (if relevant)

Although the parameters could have been regularized using Ridge or ElasticNet, – as opposed to Ridge regularization. ElasticNet is a combination of Ridge and Lasso and would therefore be a compromise between the two (Hastie et al., 2009). The shrinking of parameter estimates to zero gives a smaller number of features. This has the benefit of reducing the probability of a spurious feature-target correlation that would result in an overfit ML model (Hawkins, 2004).

Given that this replication did not process the same data, nor used the same techniques for neither partitioning, feature scaling, feature selection, or for machine learning model, it is not surprising that the results differ (see table x \* for short summary).

The data acquisition step varied greatly as there were dissimilarities in the participant pool, the task and in both the length and number of recordings.  
It can be hypothesized that conditions such as alogia or the flat effect sometimes found in patients that are thought to elicit some of the acoustic atypicalities might manifest itself differently across languages. The fact that this replication had participants speak Danish as opposed to English might impact the ML algorithms ability to detect patterns for classification. Moreover, none of the participants spoke their first language in the original study given their Malay, Indian or Chinese origin. As of yet, research points towards some general differences in acoustic patterns in schizophrenia patients related to symptoms such as alogia and flat the effect\* Cite \*. However, from the knowledge of this researcher, very little research sheds light on the potential modulation that language or language nativeness might induce. Moreover, the pool of schizophrenic participants might also vary between the original and this replication as people diagnosed with schizophrenia elicit slightly different symptoms (*Lundbeck Institute Campus*, 2016; Sartorius et al., 1986).

The number of recordings was significantly higher 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 more feature vectors (1 per recording) in this replication. In machine learning, each feature vector represents a data point and thus the classification algorithm simply had more datapoints to learn from. The recordings were however, substantially longer in the Chakraborty et al. study which meant that the feature vectors for each data point were more accurate and less prone to random variation \* cite \*.

As using SVM as an algorithm requires scaled parameters/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). 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 be slightly better, but it would reflect out-of-sample performance as accurately. 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).

LASSO regularization was utilized for feature selection in this study. Contrastingly, Chakraborty et al. utilized Principal Component Analysis (PCA). PCA reduces the dimensionality (number of features) of each data point (each recording), by generating a smaller number of new ‘principal components (dimensions) while preserving as much as the data’s variation as possible (Abdi & Williams, 2010). The latter feature selection technique diminishes the interpretability of the model, given that the original acoustic features are convoluted in the new principal components. Choosing one specific feature selection technique over another should in theory not have a large impact on performance in classification. Much theory supports the choice being arbitrary, but in practice it sometimes is not (Oreski et al., 2017). However, both PCA and LASSO have been found as some of the best feature selection techniques, with great improvements of classification algorithms (Sun et al., 2019). It is therefore 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 work.

1. Replication
   1. Possible
   2. Hard (Methods explained in condensed manner)
2. Comparison
   1. Possible
   2. Hard (More information on sexes and nationalities needed)
3. Getting similar results
   1. Differences in performance – where does it come from?
      1. Biased labels
      2. Difference in language
      3. Task differences
      4. Difference in algorithms
      5. Arbitrary choices for tuning
      6. A mixture (which mixture?) of all the above
   2. Some things might balance each other’s out, some might not

### 4.2.2 Problems established from conservative replication

**Actual:**

**Notes:**

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

To underline how much more shit the predictions will get when you’re rigorous  
With increased conservatism, how good are the results really?

Results real good if overfit

To underline how much more shit the predictions will get when you’re rigorous  
With increased conservatism, how good are the results really?

**Discussion:**

Many choices have to do with handycrafts and arbitrary choices (tuning)

We haven’t gotten enough research on principles of how to do this (Right now up to individual experience of researchers)

Present the issue that this hasn’t been fixed by our paper either. Even with relatively simple algorithms. Deep learning would mean that this is even worse.

So is language/way of speaking

* Hard to know whether differences in performance are due to:
  + biased labels
  + difference in language
  + task differences
  + difference in machine learning algorithm
  + arbitrary choices for tuning
  + a mixture of all (could balance each other out, if in opposite directions (some make it harder for the Danish corpus, some make it easier)

## 4.3 Further research

### 4.3.1 Need for a widely applicable, conservative, transparent pipeline.

1. Meta (widely applicable, conservative, transparent pipeline) would help by:
   1. Avoiding in overfitting (as mentioned previously)
   2. Making it easier to compare results (as mentioned previously)
      1. Within or across sexes and nationalities,
   3. Making it easier to replicate
   4. Enabling research to know locate the origin of differences in results (as mentioned previously)
      1. Biased labels
      2. Difference in language
      3. Task differences
      4. Difference in algorithms
      5. Arbitrary choices for tuning
      6. A mixture (which mixture?) of all the above
   5. Shedding light on arbitrary choices – either by:
      1. Providing information on it in the papers
      2. Providing a method for making these choices
2. In general: More replications and a generally more open-science based approach

This study was not enough.

**Discussion:**

Many choices have to do with handycrafts and arbitrary choices (tuning)

We haven’t gotten enough research on principles of how to do this (Right now up to individual experience of researchers)

Present the issue that this hasn’t been fixed by our paper either. Even with relatively simple algorithms. Deep learning would mean that this is even worse.

1. Establish a widely accepted pipeline. It should be:
   1. Rigorous
   2. Conservative
   3. Transparent
2. It allows for
   1. Replications
   2. Transparency

To test generalizability and robustness

This allows for more replications

# 5. Conclusion

**Actual paper:**

**Notes:**

# 6. Acknowledgements

s

# 7. References

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

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.

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

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

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

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

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

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

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.

# 8. Appendix

s