AARHUS UNIVERSITY

VOICE ATYPICALITIES IN SCHIZOPHRENIA; REPLICABILITY OF MACHINE LEARNING APPROACHES

Author:
Emil Trenckner Jessen - student
(au604547@uni.au.dk)

Supervisor:
Riccardo Fusaroli - Associate professor

University:
Cognitive Science, Aarhus University
Jens Chr. Skous Vej 2,
8000 Aarhus, Denmark

Contents

Abstract

Can machine learning (ML) applied to voice data be used as a tool to help diagnose and track individuals 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. Research suggests that the field suffers from problems of bias, overfitting, and models with low robustness and generalizability. This study provides a proposal for a conservative machine learning pipeline suitable for reducing these problems. As a way of exemplifying its use as well as facilitating replications it is furthermore used to replicate of the promising study by Chakraborty et al. (2018). The replication resulting in a macro average F1-score of 0.70 - 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 pipeline's capacity for alleviating problems within the field was evaluated. Across-sample testing and open-science conduct was proposed as a way to access information about generalizability and robustness of the method.

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 differences (Andreasen et al., 1995; Cohen et al., 2012; Covington et al., 2005; Kuperberg, 2010; Parola et al., 2019). These 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). The prospects of using

speech clinically are ample but impractical on a larger scale. As a result, recent endeavors have been trying to automate the assessment using supervised machine learning (ML) approaches.

1.2 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 features and class labels (Kotsiantis et al., 2007). After training a model, it can then be used to classify 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 cheaply and give clinicians an effortless way of tracking and predicting progressions in symptoms. Furthermore, judgements would be 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; Parola et al., 2019; 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.

1.3 Limitations of the current literature

Although the method of machine learning looks promising at first glance, some substantial hurdles in the way of instantiating these computational methods clinically.

One hurdle is the issue of overfitting 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).

Another hurdle is the potential bias of the models. A large discrepancy of results has been found across studies within the field. This undermines belief in the generalizability of the models (Hitczenko et al., 2020). Much literature has not controlled for sociodemographic factors such as age, education,

sex and race and as a result, have produced biased models that fail to generalize to new data (Vabalas et al., 2019).

A final obstacle 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), it can be hard to investigate why. 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 the extent which they document methods and results. Given inadequate information, it is hard to pinpoint which factors cause what (Hitczenko et al., 2020).

1.3 Alleviating the barriers of ML

1.3.1 Through replications and conservative ML implementation

Replications and conservative ML implementation might prove to diminish the limitations. 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 it ensures:

a) replicability – studies must be transparent and properly document the entire process of conducting the study

b) 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 thus ensure that the two previously mentioned criteria are met. But what constitutes a conservative ML implementation? This paper will attempt to provide a general pipeline that may guide good ML conduct. The workflow that the pipeline suggests will - if implemented - allow for unbiased models as well as improve the conditions for comparisons of methods and results across studies.

1.3.2 A general pipeline for ML using voice

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

The pipeline that this paper is presenting is broad and general, with aspirations of being widely inclusive. Its intended use is within ML research using voice as a predictor and may be directly applicable to research within the fields of autism or depression. The pipeline will narrow in the range of options to ensure that the necessary requirements for conservative ML conduct are being met. The pipeline will be divided up into 9 steps, which can be seen visualized in figure 1. The pipeline will not specify exactly how these steps ought to be carried out. Proper and transparent documentation is therefore critical – just as in all research, but perhaps especially within a field that suffers from little replicability and poor documentation (Vabalas et al., 2019).

1) Data acquisition. A thorough understanding of the data is important for avoiding pitfalls. A number of factors from data can confound a study if precautionary measures are neglected. First, it is important to be wary of any bias that might arise in the model as a result of sociodemographic factors. Educational level, age, race, sex have been known to cause a wide array of harmful bias across research fields, but additional factors such as medication and severity of symptoms might also contribute to biases (Blodgett et al., 2020; Cohen et al., 2016; Hitczenko et al., 2020). Secondly, data quantity is important. Internal and external validity of a study have been found to undermined by small sample sizes (Faber & Fonseca, 2014). An association between small sample sizes and biased performance in ML studies classifying diagnosis from voice have also been found (Vabalas et al., 2019). Thirdly, the task from which the recordings are derived must be considered. Cognitive and social load has been found to increase the effects of schizophrenia in the acoustic signal (Parola et al., 2019). Predicting recordings from a harder task might therefore elicit good results, that do not generalize to easier task recordings. Finally, irrelevant recording identifiers must be controlled for. Background noise, room ambience or recording settings should ideally be uniform in the data. Having all schizophrenics all be recorded within one room and the healthy controls in another could cause potential problems as acoustic features of participants might be altered by room acoustics (Olsen, 2018).

2) Preprocessing. Preprocessing includes noise removal and potential data augmentation. This step may alleviate the data of confounds such as room acoustics or differences in microphone settings (Olsen, 2018). Since raw recordings cannot be used to predict, features also must be extracted from the speech within this step. The choice of features set can be driven by theory, by choices of past studies or can be entirely explorative.

3) Data partitioning. Train-test splits have found to be more robust and provide less balanced results in comparison to K-fold cross-validation and can be beneficial (Vabalas et al., 2019). Further division of the training set, into a training and a validation set, can allow better hyperparameter tuning (Schratz et al., 2019). The ratio of train-test splits has an impact. Larger training sets allow for a model to better learn the patterns in the data, while a larger test set allows for a more accurate measure of performance. Having for example only three voice recordings to in the test set could only result in an accuracy of either 0, 33.33, 66.67 or 100 percent accuracy even given a true accuracy 70 percent. Although there is some basis for choosing the split, there is no scientific consensus on what is optimal. 80/20 is often used (Hastie et al., 2009). Given an unbalanced dataset, some precautionary measures ought to be taken when splitting. An unbalanced training set of for example 4 male patients and 2 female controls, might lead to the model predicting 'schizophrenic' to all cases where the acoustic features are specific to males. A model might end up biased, if it learns the acoustic patterns of males instead of those for schizophrenia (Leavy, 2018). Given testing on a set with many females and males, this bias can be investigated, however.

4) 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 (Hastie et al., 2009). Regardless of scaling method, 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 test set, which otherwise would result in overfitting (Myrianthous, 2020).

5) Feature selection: It can be necessary to select a subset of features, given many extracted features. Feature selection is carried out in order to improve predictive power and interpretability as well as to reduce complexity and need for computational power (Hastie et al., 2009).

Features must only be selected on basis of information in the training set, and not on the pooled training and testing data. Selecting relevant features based on what is relevant in the test set is going

to produce problems with 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. They do in theory perform the same task, but 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).

6), 7), 8) Model training, tuning and validation. Supervised machine learning covers a wide range of algorithms that all produce models based on some set of data. 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 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). Optimal values can, however, be discovered semi-manually. One of the benefits of partitioning the data up into a training, a validation and a test 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, its performance can be explored via the validation set. The hyperparameters can then again be tuned and repeatedly be validated until the optimal hyperparameter settings have been found. Since the model has been validated without the use of the test set, the model has not been overfit to the test set, thus making it suitable for evaluation of true performance.

9) Test and evaluation. When evaluating performance on the test set, confusion matrices are critical. They provide the complete picture of performance and all relevant metrics of performance can be calculated solely using the information from the matrix. They ought to be supplemented by additional evaluation metrics, however. Accuracy – the percentage of correct classifications – is regarded common practice but can often misleading, which is why other measures such as precision, recall and F1-scores ought to be provided (Hossin & Sulaiman, 2015).

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 appear promising in distinguishing and schizophrenia. However, the field of machine learning within this topic have issues with overfitting,

bias, and with comparability of results between studies - a result of large differences in methods between studies.

To alleviate these problems, this study provides a pipeline which assists in diminishing issues of overfitting and bias as well as improving conditions for comparison of results between studies. As a way of facilitating replications and providing an exemplification of the pipeline, this study will furthermore perform a replication of the study by Chakraborty and colleagues from 2018 (Chakraborty et al., 2018).

2. Methods

2.1 Pipeline implementation

The replication of this paper follows and provides an exemplification of the proposed pipeline. The methods section will provide a detailed description of the choices for each step. An overview that showcases how this study made use of the proposed pipeline can be seen in figure 1, below. Source code can be accessed via https://github.com/emiltj/bachelors. Preprocessing, partitioning, and feature selection work was carried using using R, Rstudio, (R Core Team, 2019; RStudio Team, 2020), while feature extraction and ML modeling utilized openSMILE and Python, respectively (Eyben et al., 2010; Van Rossum & Drake, 2009)

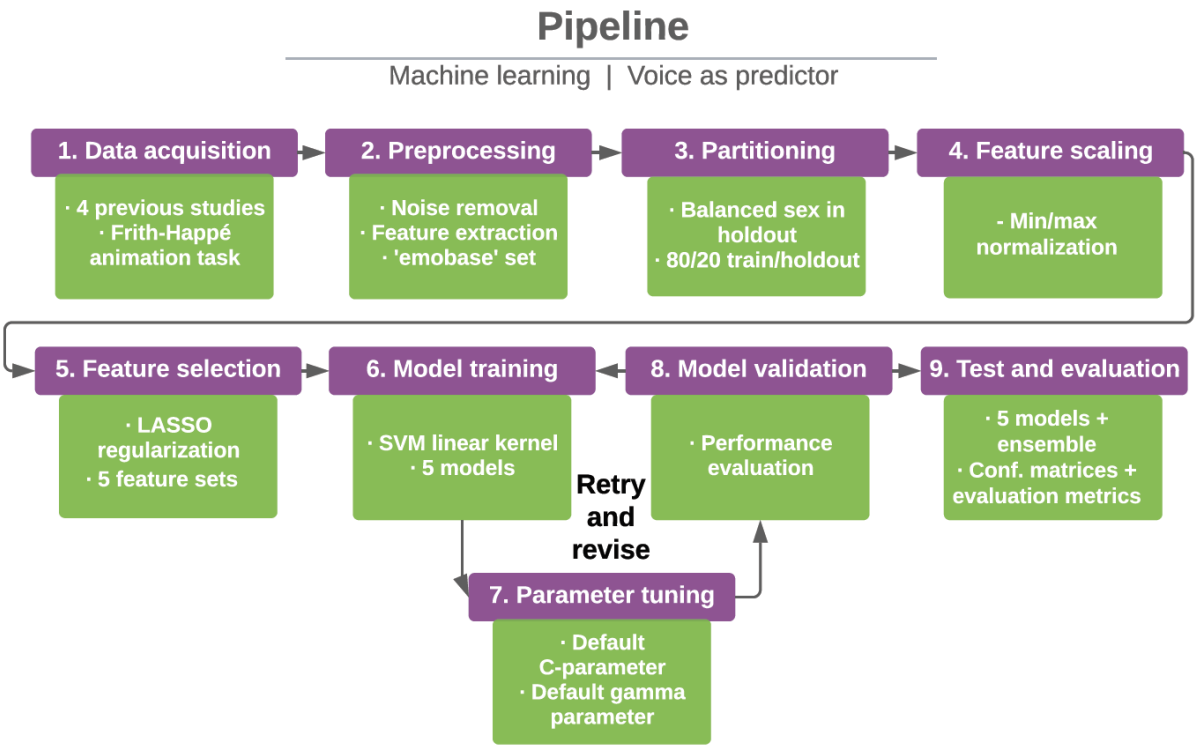

Figure 1.

An overview of the proposed pipeline. Purple boxes refer to the general pipeline whereas the green refer to the specific choices of this replication.

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. (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 forward to Sep 15, 2020 when the continued search took place. The continuation of the search used 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 title. Relevant papers, 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 large amounts of data. This narrowed the number of papers down

to 8 papers (see appendix, 7.1). The study by Chakraborty and colleagues was chosen for replication after carefully assessment. (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.

Participants from all studies went through the same task; namely the Frith Happé animations task (Abell et al., 2000). The task consisted of watching 2D top-view videos of animated triangles moving around on the screen. After watching an animation, the participants were interviewed and asked to describe what happened in the animation. 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 that (mean duration = 18.18 sec., SD duration = 14.84). Recording settings and equipment was constant within study, but unique across studies.

2.3.2 Participants

222 Danish participants were included in this study. Out of the 222 participants 106 were clinically diagnosed with schizophrenia by the standards of ICD-10 DCR (Zivetz, 1992). Patients were recruited through OPUS, 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 and educational level. Healthy control subjects were recruited via advertisements in four local newspapers. The control group (and their first-degree relatives) had no history of psychological disorders. 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 within each of the original studies. N, SD, HC, SZ refers to number, standard deviation, healthy controls and the schizophrenia group respectively.

2.4 Preprocessing

2.4.1 Cleaning of audio files

The cleaning of the audio files was carried out by Ludvig Olsen (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 (Eyben et al., 2010). The 'emobase' base-set configuration file of 988 emotion recognition features was used to extract features from the recordings. The 'emobase' feature set contained 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 (for full list of features, see appendix *). 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 test set consisting of 80% and 20% of the 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 test set. Moreover, the test set was forced balanced – both on the account of sex and diagnosis. The test set contained feature vectors for each trial from 23 controls (11 female) and 21 patients (10 female).

The training data was then furthermore divided up into 5 folds – similarly keeping recordings from the same ID within the same fold. These folds were used to create 5 training sets and 5 validations sets. Each training set consisted of 4/5th of the full training data, while the validation sets consisted of the remaining fold.

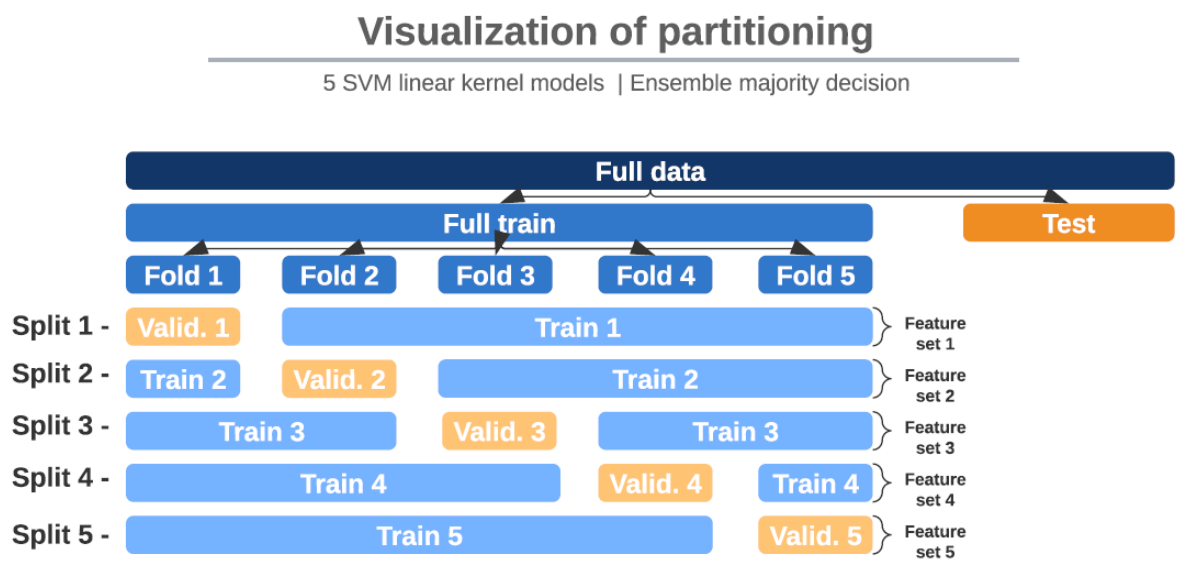

Figure 2.

A visualization of the multi-leveled partitioning of the data. The data was first split into a training and a test set and the full training data was then split into 5 folds. These folds were subsequently used in 5 sets of training and validation sets. LASSO feature selection was performed on each of the 5 training sets, resulting in an appertaining feature set for each training set.

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. Normalization was carried out separately for the full training data and the test data – both using the min-max values from the full training set.

2.7 Feature selection

Feature selection was carried out using the Least Absolute Shrinkage and Selection Operator (LASSO) analysis regression. The R package 'glmnet' was utilized for the purpose of this paper. (Friedman et al., 2010). LASSO optimizes beta estimates for all features through a loss function based on misclassification error and an added regularization term.

For this paper, the latter term utilized what is known as 'lambda.1se' - the lambda value resulting in the fewest number of features within 1 SE of the lambda value that minimized the loss function. As the full training data had been divided up into 5 splits (see fig. 2), LASSO was performed on the 5 training sets separately which resulted in a feature set for each (for list of features in the sets, see appendix x*). An illustration of the feature selection for one of these splits can be seen below in figure 3.

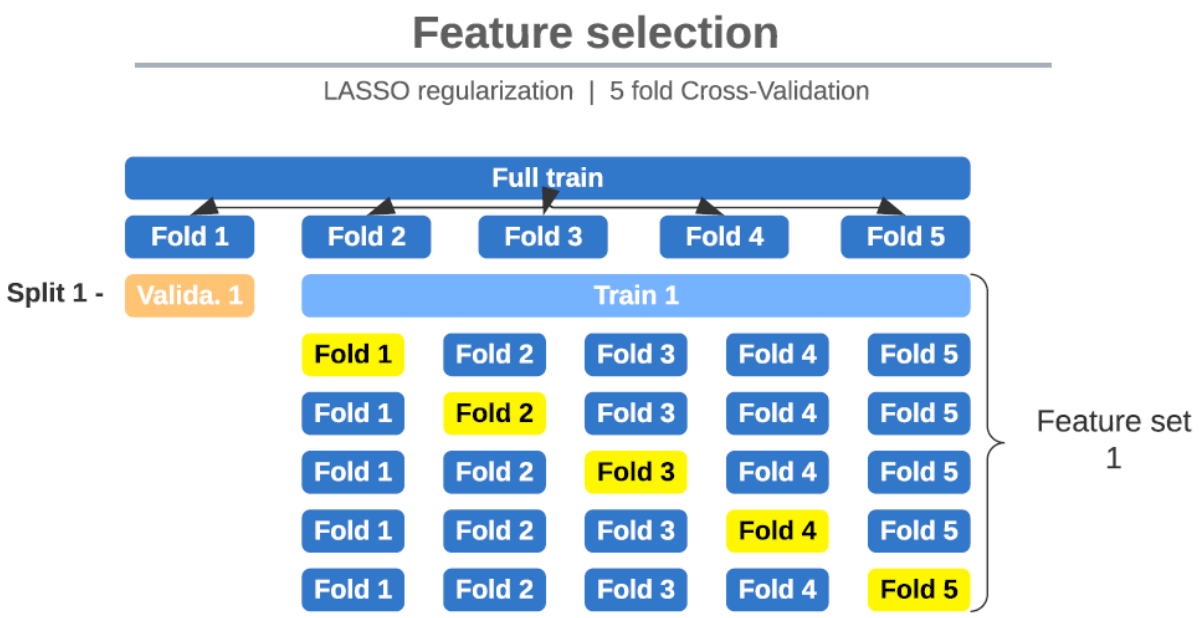

Figure 3:

Figure showing the process of feature selection on the first training set. Train 1 is divided up into 5 folds. A single fold is then excluded (yellow). Using cross-validation, the LASSO regression fit a range of lambda values, to find the optimal value for classification on the excluded fold. The features selected using the optimal value thus constituted the feature set. This entire procedure was then repeated for the remaining 4 splits.

2.8 Model training, tuning and validation

Using the 5 training sets and the appertaining feature sets for each split, 5 SVM linear kernel classifier models were constructed using the Scikit-learn module in Python (Pedregosa et al., 2011; Van Rossum & Drake, 2009). The SVM classifiers were then validated on the appropriate validation sets, repeatedly using a range of C-parameters. Performance was tracked using the metrics specified in section 2.7. The default C-parameter of 1 was found optimal for classification and was used for training all 5 models before the final testing.

2.9 Test and evaluation

2.9.1 Testing the models

The 5 models were then assigned to distinguish schizophrenia voice recordings from the test set. An ensemble model was furthermore created. This ensemble model used the previously mentioned models as constituents and its predictions was the majority vote. If for example 3 out of 5 models predicted 'schizophrenia' for a recording, then this was also the vote of the ensemble model. Performance on the testing set was then evaluated for the 6 models using relevant metrics. To allow for insights into within-sample performance, the 5 models were also assigned to predict the recordings that the models had learned from (training set 1-5). They also predicted the recordings from the validation sets.

2.9.2 Evaluation metrics

For evaluating the performance of the models, several metrics conveying information about the classification was provided. Information on precision, recall and F1-scores for each class (controls and patients) was provided, along with a macro average F1-score, accuracy and baseline accuracy. Moreover, confusion matrices are provided as they convey the whole picture of performance and provide all the information needed for calculations of all evaluation metrics.

$$precision = \frac{tp}{tp + fp}$$

$$recall = \frac{tp}{tp + fn}$$

$$f1 = \frac{2 * precision * recall}{precision + recall}$$

$$macrof1 = \frac{1}{N}\Sigma_{i=1}^{N} f1_i$$

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn}100$$

$$baselineAcc = \frac{max(tp + fn, fp + tn)}{tp + fn + fp + tn}100$$

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 meant that it diverged from the original study on several aspects. The discrepancies can all be seen in table 2, below.

	Chakraborty et al. (2018)	**Replication**
N (participants)	78	222
Female rate	52.6%	42.8%
SZ rate	66.67%	48.2%
Origin	Malay, Indian, Chinese	Danish
Task language	English	Danish
N (recordings)	78 (1 per participant)	1900 (8-10 per participant)
Mean (recording length)	26 min	18.8 sec
Feature selection	PCA	LASSO regularization
Feature scaling	Min-max normalization	No information
ML algorithm	Single SVM	Majority vote ensemble – SVM
Final testing set	Cross- validation (full dataset)	Test set (separate set for final test)

Table 2:

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 3. An in-depth look at the ensemble models performance; both for controls and for the patient group is provided in table 4. Table 4 also provides insight into performance differences between the sexes. Finally, a confusion matrix (table 5) provide the full picture of performance. The latter two tables use 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
Validation 1	Train 1	0.687	68.68%	51.85
Validation 2	Train 2	0.630	63.05%	54.34
Validation 3	Train 3	0.678	67.84%	51.62
Validation 4	Train 4	0.613	61.31%	52.94
Validation 5	Train 5	0.658	65.80%	53.29
Test	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 3:

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

Within-sample prediction performance can be seen in row 1-5, while row 5-10 depicts performance

tested on the 5 validation sets. Finally, the performance for the models' predictions on the test 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
Test	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 4:

Performance of the ensemble model - within both each sex and within HC/SZ.

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 5:

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. 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 pipeline work out in this replication?* ", 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 gained from this replication. Future research using the pipeline will also be discussed, looking into both benefits, limitations and development.

4.1 Model performance

This section will compare performance of the original paper with the performance of the ensemble model. Performance on the test set will be investigated as this is what gives information about the out-of-sample capabilities of the model - as opposed to looking at the predictions on the training or validation set which would not give an idea of the generalizability of the model.

The original paper had a macro average F1-score of 0.77 – higher than that of this replication (0.70). When looking at the isolated F1-scores for classifying patients and controls, both models classified controls equally well. The model from the original study did, however, achieve a higher F1-score when classifying patients (0.84) compared to this study (0.70). 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, the models may have elicited biases. 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 (Chakraborty et al., 2018).

All performance measures considered, a moderate difference in performance was found with this replication seemingly having slightly worse classification capabilities. This can be interpreted in various ways. Was it due to the differences in data? Or was it due to not applying the optimal methods in this replication? To shed light on the difference in performance the specifics of the pipeline implementation and their divergence from the original study will be evaluated.

4.2 An evaluation of the pipeline implementation

The proposed general pipeline did not provide a rigid guide to the specific execution; therefore, the specific choices must be evaluated. The impact of the deviation between the studies will also be discussed in relation to the differences in performance (for an overview of deviations, see table 2).

1) Data acquisition. This study used data corpora of diverse speech recordings from multiple studies. This meant that the data was more diverse data since the recording setting differed across study. The ML model is therefore likely to be slightly more robust, in that it is less bound to only learning patterns within a certain setting.

All participants in this study were Danish while the original study employed Malay, Indian and Chinese participants. This means that the data differs from the original study. Culture has been known to modulate symptoms of schizophrenia – with for example westerners eliciting stronger depressive behavior (*Lundbeck Institute Campus*, 2016; Sartorius et al., 1986). Moreover, sociodemographic factors have been known to play a role in the acoustic differences as well (Hitczenko et al., 2020). The original study furthermore instructed their participants to speaking English during their recordings– a non-native language. As cognitive load has been found to show larger symptomatologic effects for voice in patients (Parola et al., 2019), this might have elicited stronger patterns for the model to pick up and correspondingly better predictions.

The number of recordings for this study (N = 1900) was significantly greater than the original study (N = 78), as a result of having 8-10 recordings for each participant in the replication. Since each recording only produced a single datapoint, the algorithm had more data to learn from in this replication. The recordings in the study by Chakraborty et al., did however have much longer recordings, which meant that for each data point, the true features values were more accurately captured.

2) Preprocessing. For feature extraction, the 'emobase' feature set was utilized to capture the acoustics of the emotional impairment of schizophrenia (Eyben et al., 2010). However, many other

feature sets could have been used. It would have been interesting to use multiple feature sets – such as the features from DigiVoice, either in conjunction or for comparison (Zhang et al., 2019).

3) Partitioning. This study was roughly balanced on sex and diagnosis. The holdout set included enough male (N = 194) and female recordings (N = 180) to allow for insights into whether the slightly fewer numbers of females in the training data confounded the results. As discussed in 4.1.1 the model was unbiased in terms of sex. The original study was balanced but offered no information on potential bias.

4) Feature scaling. Feature were scaled using a min-max normalization. Alternatively, standardization could have been utilized to be less affected by outliers, since standardization uses standard deviation as opposed to max-min values. The scaling of both the training and holdout set solely used information from the training set to avoid overfitting (Géron, 2019; Myrianthous, 2020; Vabalas et al., 2019). As no information was provided in the original paper, it is unclear whether they scaled similarly. They might have scaled prior to the cross-validation that they used – allowing for overfitting - or instead within each step of the cross-validation. Given the former option performance would appear better, but it would more poorly reflect out-of-sample performance.

5) Feature selection. Feature selection was in both the original and in this replication carried out using only information from the training set which avoided overfitting – a measure often neglected within this field (Vabalas et al., 2019). LASSO regularization and Principal Component Analysis (PCA) which was utilized in this replication and the original study respectively have been found to be similar in performance, with great improvements of classification algorithms (Abdi & Williams, 2010; Sun et al., 2019). Given that they perform similarly, it is unlikely that all variation in performance between the studies can be attributed 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. For this paper, it could have been informative to explore PCA and the many other techniques and compare their performance to shed light on this topic.

6, 7, 8) Model training, tuning and validation. SVM linear kernel models were utilized in both the replication and the original study. However, the use of an ensemble model was 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 test set this speculation would require further testing across datasets.

9) Testing. Testing of the models was similarly carried out in this replication and the original. Discussion of result has been provided in section 4.1. Supplementary metrics could have been provided. ROC-curves show the tradeoff between false positives and false negatives at different classification thresholds – knowledge that is important to know before applying such models clinically. Given that research has not yet established the generalizability and ecological validity of these ML algorithms, they were deemed unnecessary and thus omitted.

In summary; this replication is unlikely to have been confounded by problems related to data, bias on the basis of sex or overfitting. 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 how and to which magnitude since other factors were not controlled for.

4.3 Limitations and prospects of the pipeline

By providing a general pipeline for classification of schizophrenia patients, it is the 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 and bias within the literature. However, the proposed pipeline is not an exhaustive solution. 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 inclusive and broadly applicable. However, the generalist nature of the pipeline has a downside. Many of the choices for good ML conduct are still up in the air, which hosts room for error. Choices for algorithms and feature selection technique are still left up to the practical experience of the individual researcher. The problem of knowing which algorithm to use for instance is not aided by the proposed pipeline. 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 ML implementation. Using this replication as a case example, it is apparent that it proves difficult to pinpoint what drove the performance

differences. It could be attributed to the difference in data – participants came from different backgrounds; they spoke different languages and were also presented to task of dissimilar nature. However, overfitting could have played a minor role, since the original did not document whether feature scaling was performed on the combined training and testing set or not.

The future use of testing across datasets combined with the use of a rigorous pipelines could potentially help solve the limitations within the field (Hitczenko et al., 2020; Vijayakumar & Cheung, 2018). Testing across datasets would shed light on generalizable out-of-sample performance. Combining this with a pipeline could contribute in streamlining research and making comparisons of models easier. However, data and models from other researchers can be challenging to retrieve – data and models are often lost over time and often include sensitive information. Data sharing may seem unfeasible, but studies within similar research areas have proven that these hurdles can be overcome by actively anonymizing and storing data (Gratch et al., 2014; Schuller et al., 2013). Such initiatives within voice in schizophrenia would be beneficial, not only for knowledge of generalizability of ML research, but also for pinning down which acoustic features differentiate healthy individuals from individuals suffering from schizophrenia (Parola et al., 2019). This paper therefore suggests that research should strive towards employing more open-science conduct – sharing of data, scripts and models.

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

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

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM International Conference on Multimedia*, 1459–1462.

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.

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%20i

t%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

Gratch, J., Artstein, R., Lucas, G. M., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., & Marsella, S. (2014). The distress analysis interview corpus of human and computer interviews. *LREC*, 3123–3128.

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

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

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

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 1.

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*.

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

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

Leavy, S. (2018). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering*, 14–16.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

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

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

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

RStudio Team. (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC. http://www.rstudio.com/

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

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., & Marchi, E. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.

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

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

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

Vijayakumar, R., & Cheung, M. W.-L. (2018). Replicability of Machine Learning Models in the Social Sciences. *Zeitschrift Für Psychologie*, *226*(4), 259–273. https://doi.org/10.1027/2151-2604/a000344

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 'Emobase' feature set

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.

6.3 Feature lists after L2 Regularization

https://github.com/emiltj/bachelors/tree/master/final_feature_sets