Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5B
D-70569 Stuttgart

**Master's Thesis**
**M.Sc. Computational Linguistics**

# Creating a robust and effective feature selection pipeline in the clinical setting

## How to leverage information from multiple modalities to identify features that are health condition-sensitive and -specific?

**Vanessa Richter**

| | |
|---|---|
| *Examiner:* | Prof. Dr. Ngoc Thang Vu<br>Dr. Antje Schweitzer |
| *Supervisor:* | Florian Lux |
| *Start:* | 13.01.2023 |
| *End:* | 13.07.2023 |

**Erklärung (Statement of Authorship)**

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig verfasst habe und dabei keine andere als die angegebene Literatur verwendet habe. Alle Zitate und sinngemäßen Entlehnungen sind als solche unter genauer Angabe der Quelle gekennzeichnet. Die eingereichte Arbeit ist weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens gewesen. Sie ist weder vollständig noch in Teilen bereits veröffentlicht. Die beigefügte elektronische Version stimmt mit dem Druckexemplar überein. [1]

(Vanessa Richter)

[1] Non-binding translation for convenience: This text is the result of my own work, and any material from published or unpublished work of others which is used either verbatim or indirectly in the text is credited to the author including details about the exact source in the text. This work has not been part of any other previous examination, neither completely nor in parts. It has neither completely nor partially been published before. The submitted electronic version is identical to this print version.

# Contents

**Abstract**

Utilizing computer vision and speech signal processing to assess neurological and psychiatric conditions has the potential to help detecting diseases or monitoring their progression earlier and more accurately. However, retrieving the required information from speech and facial modalities presents the challenge of finding features that generalize across studies with high sensitivity and specificity. A major task in finding such features is dealing with overfitting to data biases in small sample sizes and redundancy in the analysis of high-dimensional feature sets. It is also critical to ensure interpretability of these methods since the results of health screening tools must be explainable to clinicians and patients.

In this thesis, we present a transparent feature selection pipeline that specifically addresses demographic biases and feature redundancy. Our method provides interpretable insights by quantifying feature contributions to classification results using Shapley values. More specifically, we assessed age trends of the entire healthy control cohort and corrected the feature values based on the determined age coefficients. Sex-specific z-scoring was used to account for differences between males and females. To address feature redundancy, we used hierarchical clustering to group features into sensible domain-specific clusters, such as voice quality, jaw movement, or mouth symmetry. These clusters together with feature effect sizes were used in the classification step to select only the most salient features as input to the classifier. Finally, Shapley values were calculated to unwrap model decisions and evaluate the contribution of individual features.

We used datasets on neurological (bulbar pre-symptomatic and bulbar symptomatic ALS) and mental (depression and schizophrenia) diseases as well as a healthy control dataset. The data was collected in a real-world scenario, where participants engaged with a virtual agent that guided the participants through a set of tasks.

We apply the presented feature selection method including Shapley-based analyses on these datasets. Our analysis provides valuable insights into feature contribution among binary and multiclass classification experiments and reveals shared characteristics across disorders.

# 1 Introduction

## 1.1 Problem Description & Motivation

One out of eight individuals in the world lives with a mental health disorder, but most people do not have access to effective care [2]. Moreover, disorders of the nervous system are the second leading cause of death globally (Feigin et al., 2019). The development of clinically valid digital markers for neurological and mental disorders that can be extracted automatically could help improve patients' lives by releasing the capacity of clinicians and physicians and providing an objective and faster assessment than standard clinical judgments. By that, it aids early diagnosis while making health care more accessible and affordable. This poses a great opportunity for improving patient care.

Leveraging information from multiple signal modalities (such as speech, natural language, motor-behavior and video) to distinguish patients from healthy controls or to monitor progression in a particular disease comes with the challenge of finding features that generalize across studies and that are not only sensitive but also specific to a disease. A major challenge in finding such features is dealing with overfitting as well as redundancy when analyzing high-dimensional feature sets. Furthermore, an exhaustive search of all possible feature combinations to find the subset maximizing the classification performance is not feasible when it comes to feature sets containing hundreds or thousands of potential markers.

Self-supervised learning has been found an effective method to achieve excellent results with raw audio data without requiring costly feature engineering or a large amount of labeled data (Baevski et al., 2020; Hsu et al., 2021). In the area of speech recognition, Baevski et al. (2020) demonstrated that latent speech representations learned from speech audio alone and a subsequent fine-tuning on transcribed speech can outperform the best semi-supervised methods. At the same time, it is conceptually simpler.

However, in the clinical setting, it is critical to ensure interpretability of methods, as the results of health screening tools should be explainable to physicians, clinicians, and patients. Learned feature representations do not serve this purpose. In general, classification models are a useful tool to assess the discriminatory power of the features under investigation, but maximizing their performance is not an isolated aim. The main objective is to identify multimodal biomarkers for a disease, which is only possible through careful, manual and transparent feature development. This is related to the fact that the selected features should be clinically meaningful.

The performance of Machine Learning (ML) models typically increases with the size of the

[2] https://www.who.int/news-room/fact-sheets/detail/mental-disorders, accessed 11/7/2022

dataset. However, Berisha et al. (2022) found a counter-intuitive negative association between classification accuracy and sample size in clinical speech ML literature, implying that the classification performance decreases with larger samples. As large sample sizes should help the model to generalize, their findings indicate that the good performance of models trained on small sample sizes might rather be due to model overfitting and publication bias than disorder-relevant features. More thorough research towards pairing clinically valid feature selection with adequate discrimination performance of classification needs to be done, as such models are intended to be used in a sensitive area where inaccurate results impact the lives of the individuals concerned, making the stakes extremely high in case of erroneous results. These findings further suggest that a careful clinically-supported, individually-validated feature selection may be essential for dealing with small sample sizes as such features are more likely to help the model to generalize (Berisha et al., 2021).

In relation to this, another challenge in the mental and neurological health space is that datasets are often not only small but also unbalanced (Aleem et al., 2022). A major difficulty is to distinguish between features that are predictive of a particular disease and those that are related to idiosyncrasies of the dataset. It has been found that not only speech characteristics, but also facial behavior change with age (Hunter et al., 2012; Malatesta et al., 1987) and differ by sex (Dimberg and Lundquist, 1990). Moreover, classification accuracy based on facial behavior has been found to vary between males and females (Drimalla et al., 2020). For example, what is considered a healthy characteristic in an elderly individual may indicate a disorder in a young adult. Therefore, it is important to consider these demographic variables when selecting features. In this thesis, we correct for trends in both sex and age, aim at avoiding redundancy while ensuring statistical validity, and use Shapley values to uncover model decisions in binary and multiclass classification experiments across mental and neurological disorders. To our knowledge, no study to date has examined the selection of disorder-relevant multimodal biomarkers in a real-world dataset using such a comprehensive but transparent method.

The thesis is structured as follows: In Section 2, we give a brief overview of feature selection approaches and explanation of Shapley values, which we use for the purpose of model interpretability. A summary of related work is given in Section 3. In Section 4, we provide information about the dataset creation process. This is followed by a description of the proposed pipeline in Section 5. Section 6 goes into detail about our findings on demographics, redundancy as well as classification performance and feature contributions. Section 7 highlights the most important findings and discusses those in more detail. We conclude this thesis with a summary in Section 8.

## 1.2 Ethical Considerations in Remote Health Assessments

Remote health assessments require a very careful ethical handling. It is crucial that such assessments respect the autonomy of the patient. In addition, they should also be interpretable and explainable to provide adequate information to support a physician, but not to replace their diagnoses (Batliner et al., 2022). In addition to *autonomy*, *beneficence*, *nonmaleficence* and *justice* are among the four main ethical principles that should be respected (Varkey, 2020).

**Autonomy** This principle includes the support of informed consent, truth-telling, and confidentiality (Varkey, 2020). The dialog system that we use for data collection is HIPAA compliant and the research studies are approved by Institutional Review Boards (IRB)[3] to ensure the users' privacy and data security. Autonomy is ensured by requiring users' to give their informed consent before study participation.

**Justice** In health care, most pertinent is *distributive justice* (Varkey, 2020) as it refers to the fair, equitable, and appropriate distribution of health-care resources. In this regard, remote health assessment could be a facilitating and less costly application for patient monitoring, and thus provide a better accessibility for individuals.

**Beneficence and nonmaleficence** These principles state that applications should promote patients' benefit and welfare (*beneficence*) whereas harm should be avoided (*nonmaleficence*). In this thesis, our goal is to develop a robust feature selection pipeline that helps in objectively and reliably assessing clinically valid and interpretable biomarkers that aim at supporting clinicians in making accurate, objective and potentially faster diagnoses. We avoid harm by carefully researching the clinical significance of potential biomarkers and by not providing a diagnosis, neither to patients nor clinicians. Evaluated features and classification models are used for research and to inform clinicians.

[3] IRBs use a group process to review research protocols and related materials. The aim of IRB review is to ensure the protection of rights and welfare of humans participating in the research. https://www.fda.gov/about-fda/center-drug-evaluation-and-research-cder/institutional-review-boards-irbs-and-protection-human-subjects-clinical-trials

# 2 Background

## 2.1 Feature Selection Methods

Feature selection methods can be divided into four broad categories: *filter*, *wrapper*, *embedded* and *hybrid* methods while the latter combines filter or embedded with wrapper techniques (Kaur et al., 2021). Since hybrid approaches use a combination of the three basic methods, only an overview of the latter is provided here. Each type comes with different advantages and disadvantages in terms of performance, interpretability and cost.

**Filter methods** Filter methods evaluate the relevance of a feature independent of a (classification) model. Here, features are selected based on a defined (ranking) criterion and threshold, such as the statistical relationship between the input feature and the target variable.
Filter based approaches can further be categorized as univariate or multivariate, as summarized in Pudjihartono et al. (2022). In univariate approaches such Pearson correlation, euclidean distance, or the Mann-Whitney U (MWU) test features are evaluated and selected individually while in multivariate approaches, such as mutual information feature selection (MIFS) (Battiti, 1994) or minimal-redundancy-maximal-relevance (mRMR) (Peng et al., 2005), features are considered simultaneously. While multivariate approaches are able to deal with redundancy, they are more computationally expensive than univariate approaches and thereby, less effectively scalable to high dimensional data. In contrast, univariate approaches are fast and offer a high level of interpretability as they provide a straightforward ranking of features based on their relevance to the target variable. This allows for easy identification of the most important features. The use of interpretable metrics, such as p-values, Effect Size (ES), or correlation coefficients, further improves the interpretability of univariate approaches since these metrics provide direct insight into the statistical significance or strength of the relationship between each feature and the target variable.
This makes the feature selection process transparent (Kuhn and Johnson, 2013) and even high-dimensional feature sets can be processed with low cost. However, both univariate and multivariate filtering approaches do not account for model performance. Therefore, the filtered features may not be the subset with the maximum discriminatory power to distinguish between cohorts.

**Wrapper methods** Wrapper methods perform feature selection based on model performance of the chosen classifier. Features are added (forward feature selection) or removed (backward feature selection, recursive feature elimination) in order to find the optimal combination for maximizing

the classification performance (Kuhn and Johnson, 2013). A popular wrapper-based method is Recursive Feature Elimination (RFE). This method starts with fitting a ML algorithm to the whole feature set and removes the least important features until a specified number of features is reached. After each feature removal step, the model is refitted. Features are ranked by importance based on the chosen estimator (Kuhn and Johnson, 2013). Among the most popular estimators, frequently used in the areas of natural language processing and bioinformatics, are linear Support Vector Machines (SVM) (Granitto et al., 2006; Lin et al., 2012; Bedo et al., 2006). More specifically, a SVM is trained using all features initially, while, as previously explained, subsequently eliminating the least important feature. Here, the importance of each feature is assessed based on the absolute value of their corresponding weights within the SVM model.
Wrapper methods have been shown to result in higher predictive performance compared to filter methods (Wah et al., 2018; Ghosh et al., 2020). However, since optimizing classification performance is the primary goal of wrapper methods, there is an increased risk of overfitting and the selection of features is less interpretable than a filter-based approach. In addition, computation time increases significantly with the size of a feature set and finding an optimal combination for classification is not guaranteed as an exhaustive search across all possible feature combinations is usually not feasible.

**Embedded methods** Embedded methods perform automatic feature selection when the model is trained. Random forests Random Forest (RF) are an example of an embedded feature selection classification model. Those are based on a large number of individual decision trees that form an ensemble. Each decision tree in the RF generates a class prediction and the class with the majority vote is the one that the model predicts (Ho, 1995).
Neural networks are another example of an embedded method that can also perform feature selection implicitly. The process of training a neural network involves adjusting the weights and biases that connect the network's layers. Through the optimization process, the network tunes these parameters to minimize the difference between its predicted and the true output. If a particular weight becomes close to zero during training, it indicates that the corresponding feature is less relevant or redundant for the model's predictive performance. Consequently, the neural network can effectively eliminate or deemphasize less important features by tuning their associated parameters close to zero and emphasize features.

## 2.2 Shapley Values

As interpretability is important for clinical applications and ML models are less interpretable by nature, it is crucial to make their decisions more transparent.

The concept of Shapley values allows insight into the contribution of each feature to the final decision. It has its roots in the cooperative game theory literature in the 1950s (Shapley, 1951). Originally, it was introduced to determine a player's contribution to the outcome of a game. For our scenario, a feature could be seen as a player and the model's prediction as the outcome of a game. Hence, we can interpret the Shapley value as the average marginal contribution of a feature to the model prediction.
We follow Holzinger et al. (2022) for a formal introduction. The Shapley value relies on a value function $e_S$ which is defined as the expected value of the model's output $f(x)$ when a specific ($x^*$) subset of features $x_S$, where $S$ represents the subset of feature indexes, is chosen:

$$e_S = E[f(x)|x_S = x^*] \tag{1}$$

The marginal contribution of a feature $i$ compared to a set of features $S \subseteq \{1, \ldots, p\}\backslash\{i\}$ is assessed by analyzing how adding it to the set $S$ affects the value of the function $e_S$. The Shapley value $\phi(i)$ of feature $i$ is then calculated as a weighted average of the marginal contributions for all possible subsets $S$:

$$\phi(i) = \sum_{S \subseteq \{1,\ldots,p\}\backslash\{i\}} \frac{1}{|S|!(p-1-|S|)!p!} \left(e_{S\cup\{i\}} - e_S\right) \tag{2}$$

Note that the sum of all Shapley values is the model prediction:

$$f(x^*) = e_\emptyset + \sum_i \phi(i) \tag{3}$$

A strong advantage of Shapley values comes with the fact that they are based on a solid theory as they fulfill several desired axioms such as monotonicity, symmetry and linearity. However, the concept is computationally expensive as the number of possible coalitions increases exponentially with the number of features.
SHapley Additive exPlanations (SHAP) by Lundberg and Lee (2017) is a popular XAI framework and an adaptation of Shapley values that aims at providing model-agnostic local explainability. The toolkit is provided on Github[4] and used for our analysis.

[4] https://github.com/slundberg/shap/blob/master/docs/index.rst

# 3 Related Work

Feature selection methods for dimensionality reduction of clinical features have been widely studied in recent decades (Xie and Wang, 2011; La et al., 2012; Raymer et al., 2003). However, the performance of such methods is considerably variable for different datasets. For example, Bommert et al. (2020) analyzed 22 filter methods based on 16 high-dimensional classification datasets from various domains. In terms of classification performance, they found that there is no subset of filter methods that performs better than the whole rest of the filter methods on all datasets. Other studies have developed their own methodology for feature selection. Kala et al. (2014) introduced a method called *Multi-Filtration Feature Selection (MFFS)* which consists of a four-stage procedure including feature extraction, feature subset selection, feature ranking and classification. Their method adjusts the parameter *variance coverage* and the resulting feature selection quality is evaluated in terms of maximizing the classification accuracy across several classifiers trained and evaluated on synthetic medical datasets. Their method shows promising performance and robustness across the examined datasets and classifiers. The authors stress the importance of redundancy removal as they argue that classification performance is proportional to the removal of redundant features. However, the suitability for small, real-world datasets was not investigated in this study. In addition, their main aim was to optimize for classification performance rather than clinical validity. Moreover, such feature selection methods have usually been developed and evaluated for biomedical data, which are of a different nature than speech and facial features.

Recently, several speech and facial markers extracted with a multimodal dialog system have been shown to yield statistically significant differences between patients and healthy controls in neurological (Neumann et al., 2021; Kothare et al., 2022) and mental disorders (Richter et al., 2022) as well as moderate to high test-retest-reliability as a measure of robustness within the patient cohort. Furthermore, such features have been found to be useful to discriminate between patients and healthy controls with high specificity and sensitivity (Richter et al., 2022; Kothare et al., 2022; Cummins et al., 2013). Since these studies were conducted with limited sample sizes, the models are more susceptible to being impacted by group-specific characteristics, other than the investigated health condition. To alleviate such effects, several studies on visual and acoustic biomarkers ensured that patient and healthy control cohorts are age- and sex-matched (Kothare et al., 2022; Lammert et al., 2017). Others do not reference this factor at all (Williamson et al., 2016). Jiang et al. (2017) accounted for sex differences through sex-dependent modeling (GDM), i.e., male and female subjects were modeled separately. They investigated the influence of speech types, specifically interview, picture description, and reading, along with emotions in depression classification by using the classifiers K nearest neighbors (KNN), Gaussian mixture

model (GMM), support vector machine (SVM), and an ensemble method. 170 subjects (85 healthy controls and 85 participants with depression) participated in the study. They performed principal component analysis (PCA) for dimensionality reduction of the feature vector and argue that most of their features have been verified by previous research. Feature types were weighted differently according to their contribution to the detection of depression. Their results show differences in classification performance by sex with an accuracy of 80.30% for males and 75.96% for females. While they show decent classification performance in depression detection and demonstrate that performance varies by sex, they do not account for other important demographic variables such as age. In addition, their approach requires separate modeling of males and females.

As Falahati et al. (2016) point out, two popular approaches to take age, or other confounding variables, into account are either (1) using age- and sex-matched study groups or (2) stratifying the data into more homogeneous subgroups. For the latter, groups may be divided based on age, sex or both. However, the major drawback of (1) and (2) is that the data set for the respective analyses is constrained which is not desirable, in particular when dealing with small data sets. Another approach to account for demographics is to include information as auxiliary tasks for classification. Here, attention-based multi-task learning has been shown to be effective in several related areas such as NLP (Lan et al., 2017), emotion recognition from speech signals (Li et al., 2019), or image classification (Liu et al., 2021). More specifically, Li et al. (2019) included sex prediction as an auxiliary task to the principal one of emotion classification. In a multi-task setting, they deployed a neural network consisting of stacked convolutional, BiLSTM and self-attention layers for emotion classification based on speech. Their results suggest that the inclusion of sex information helped the model to better solve the main task of emotion prediction. However, to effectively learn from multiple tasks, neural networks require a substantial amount of data, which is often not feasible to obtain in a real-world medical setting.

Very recently, Muhammad et al. (2023) explored the development of an accurate and explainable ML pipeline for early detection of Parkinson's disease using multimodal time-series data. Features were extracted from the Parkinson's Progression Markers Initiative (PPMI) real-world dataset and included participant characteristics, biospecimen, medical history, motor, and non-motor function data. Their framework aims at being accurate and explainable. To select the most informative feature sets, they used recursive XGBoost and SULOV (Synthetic Unlabeled Local Outlier Voting) methods. They applied several preprocessing steps including data cleaning and balancing samples. A set of well-known classifiers such as support vector machines, RFs and light gradient boosting machines was employed. To provide model interpretability, they utilized the SHAP framework. Additionally, the authors implement LIME and SHAPASH local explainers, to further expanding the model's interpretability. The results showed that the LGBM model performed well in both the three-class and four-class prediction tasks, especially when using the non-motor function modality.

The fusion of non-motor and motor function modalities further improved the LGBM model's performance. The consistency of these explainers was explored, which they claim resulted in accurate and explainable classifiers. They also show that the calculated SHAP values provided valuable insights into the importance of features.

# 4 Data

An overview of the dataset creation process is given in Figure 1.

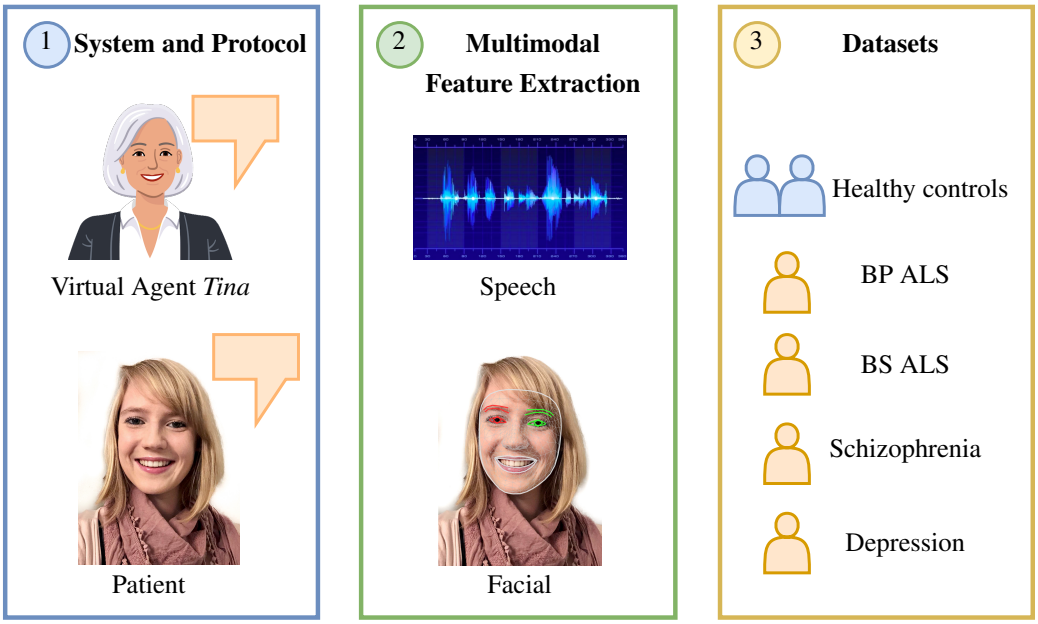



Figure 1: Overview of feature extraction and dataset creation.

## 4.1 System & Protocol

The data is collected using NEMSI (Neurological and Mental health Screening Instrument) (Suendermann-Oeft et al., 2019), a multimodal dialog system for remote health assessments. Study participants are guided by the virtual agent *Tina* through various tasks that are designed to elicit speech, facial, and motor behaviors. Having a virtual agent to elicit participants' behavior allows for scalability while providing a natural but controlled and objective interview environment and data collection. Users are provided with a website link to the secure screening portal and login credentials by their liaison (physician, clinic, referring website or patient portal). Each session starts with a microphone, speaker, and camera check to ensure that the user has given their device the permission to access camera and microphone, is able to hear Tina's instructions and the captured signals is of adequate quality. After passing the checks, the user can start the conversation. The session starts with Tina introducing herself and setting the stage. In the following screening part of the session, she involves users in a structured conversation that consists of exercises (speaking tasks, open-ended questions, motor abilities) to elicit speech, facial and motor behaviors relevant to the type of disease being studied.

In this work, we focus on tasks that are shared across study protocols: (a) sentence intelligibility test, (b) alternating motion rate diadochokinesis (DDK), (c) read speech (Bamboo task) and (d) a picture description task. For (a) participants were asked to read individual speech intelligibility test (SIT) sentences of varying sentence lengths (5-15 words), (b) required reading a longer passage (Bamboo). To assess Diadochokinetic (DDK) skills (c), participants were asked to repeat a pattern of syllables (puh-tuh-kuh /pataka/) as fast as they can until they run out of breath and (d) prompted users to describe a scene in a picture which was shown to them on a screen. These tasks are inspired by previous work (Silbergleit et al., 1997; Tomik and Guiloff, 2010; Novotny et al., 2020).

## 4.2 Multimodal Feature Extraction

We extract features of multiple modalities as shown in Table 1 that are either directly supported by clinical findings or were found to have predictive potential such as percent pause time, fundamental frequency or general quantitative and qualitative measures of facial expressivity (see for example Pueschel et al. (1998) for schizophrenia or Gaebel and Woelwer (2004) for schizophrenia and depression).

In terms of extracting facial features, the mediapipe face mesh algorithm[5] allows us to calculate 468 facial landmarks in real-time. More specifically, MediaPipe's Face Detection is based on BlazeFace (Bazarevsky et al., 2019) and determines the (x, y)-coordinates of the face for every frame. Subsequently, facial landmarks are extracted using MediaPipe Face Mesh. We select 14 key landmarks, shown in Figure 2 to compute functionals of facial behavior. Features are normalized by dividing them by the inter-caruncular distance, which is the distance between the inner canthi of the eyes (depicted as RELC and LERC in Figure 2). In terms of between- as well as within-subject analyses when the same position relative to the camera cannot be assumed, Roesler et al. (2022) found this to be the most reliable method.

Speech metrics are computed using Praat (Boersma and Van Heuven, 2001) and comprise the domains of *energy, timing, voice quality* and *frequency*. Pairing these metrics with tasks (examples of those described in the following paragraph) results in a high-dimensional set of $> X$ features.

A complete list of facial metrics and the abbreviations used is displayed in Section 4.2.

A more detailed description of speech features is shown in Section 4.2,

[5]https://google.github.io/mediapipe/

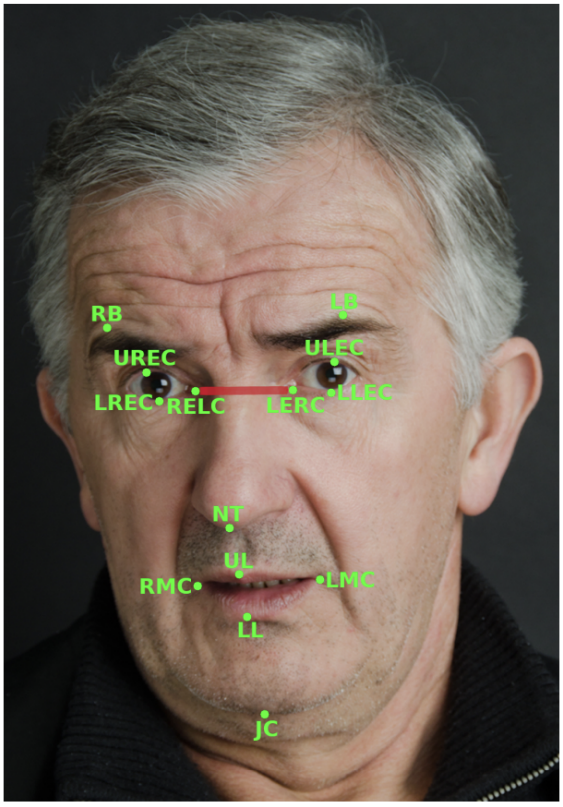



Figure 2: Illustration of the 14 facial landmarks used to calculate facial features.

| | Domain | Metrics |
|---|---|---|
| Audio | Energy | signal-to-noise ratio (SNR, dB) |
| | Timing | speaking & articulation duration/rate (sec./WPM) |
| | | percent pause time (PPT, %), canonical timing agreement (CTA, %) |
| | Specific to DDK | cycle-to-cycle temporal variability (cTV, sec.), |
| | | syllable rate (syl./sec.), number of syllables |
| | Voice quality | shimmer (%), |
| | | harmonics-to-noise ratio (HNR, dB), jitter (%) |
| | Frequency | mean, min & max fundamental frequency (F0, Hz) |
| Video | Jaw | dynamics (velocity, acceleration & jerk) |
| | Lower Lip | dynamics (velocity, acceleration & jerk) |
| | Mouth | width, opening and symmetry measurements |
| | Eye and eyebrows | opening and displacement measurements |

Table 1: Overview of speech and facial metrics.

| Metric | Description |
|---|---|
| vJC_abs_avg | mean jaw center (JC) speed |
| vJC_abs_max | max. JC speed |
| vJC_max | max. JC velocity downwards |
| vJC_min | max. JC velocity upwards |
| aJC_abs_avg | mean JC acceleration |
| aJC_abs_max | max. JC acceleration |
| aJC_max | max. JC acceleration downwards |
| aJC_min | max. JC acceleration upwards |
| jJC_abs_avg | mean JC jerk |
| jJC_abs_max | max. JC jerk |
| jJC_max | max. JC jerk downwards |
| jJC_min | max. JC jerk upwards |
| vLL_abs_avg | mean lower lip (LL) speed |
| vLL_abs_max | max. LL speed |
| vLL_max | max. LL velocity downwards |
| vLL_min | max. LL velocity upwards |
| aLL_abs_avg/max | mean & max. LL acceleration |
| aLL_max | max. LL acceleration downwards |
| aLL_min | max. LL acceleration upwards |
| jLL_abs_avg/max | mean & max. LL jerk |
| jLL_max | max. LL jerk downwards |
| jLL_min | max. LL jerk upwards |
| eye_open_avg/max | mean & max. eye opening |
| eyebrow_vpos_nt_avg/max | mean & max. eyebrow displacement |
| open_avg/max | mean & max. lip aperture |
| width_avg/max | mean & max. lip width |
| S_avg/max | mean & max. mouth surface area |
| S_ratio_avg | mean mouth symmetry ratio |

Table 2: Complete list of facial metrics and their abbreviations.

## 4.3 Datasets

For this thesis, we use control data of multiple studies and patient data on Amyotrophic Lateral Sclerosis (ALS), schizophrenia and depression of which we extract features from speech and facial modalities. Section 4.3 shows the number and age statistics of study participants for each cohort (controls, schizophrenia, Bulbar symptomatic (BS) ALS and Bulbar pre-symptomatic (BP) ALS) by sex group. As can be seen, the control cohort is with an overall mean age of 46.28 years younger than both ALS cohorts (BS: 61.48 and BP: 60.09 years), but older than the schizophrenia

| Feature | Description |
|---|---|
| SNR | Ratio of the speech signal power to the background noise. |
| Articulation duration/ rate | Duration/ rate of speech portions. |
| Speaking duration/ rate | Duration/rate of speech *and* non-speech portions. |
| PPT | Percentage of time during speech occupied by pauses. |
| CTA | Timing agreement with the speech pattern of the virtual agent. |
| cTV | Temporal variability between consecutive cycles of speech. |
| Syllable rate | Number of syllables produced per second. |
| Number of syllables | Total count of syllables in the speech signal. |
| Shimmer | Variation in amplitude of the vocal folds during the speech signal. |
| HNR | Measures the ratio of harmonics to noise in the speech signal. |
| Jitter | Variation in the timing of consecutive pitch periods. |
| Mean F0 | Average pitch of the speech signal. |
| Min. F0 | Lowest pitch observed in the speech signal. |
| Max. F0 | Highest pitch observed in the speech signal. |

Table 3: Description of speech features.

(36.46 years) and depression cohorts (34.69 years). We observe that sex groups are unbalanced for the control, schizophrenia and depression cohorts. While the schizophrenia cohort consists of more male (75.6%) than female (24.4%) participants, the control as well as the depression cohort include substantially more females than males.

An overview of the data (number of participants per sex and age statistics) used in this study is given in Section 4.3. While some datasets for a disease may be small, there is a subset of tasks that are shared across research studies. Since the data is collected in the same way (remotely with a personal electronic device), we can create a larger dataset for the healthy population across studies to get a more accurate representation of the properties of normative behavior. For the larger dataset of healthy control subjects, we identify age-related trends as well as collinearity of metric-task-combinations. This information is used to correct control as well as patient feature values from age effects and remove feature redundancies.

### 4.3.1 Schizophrenia

Schizophrenia is a chronic brain disorder that affects approximately 24 million or 1 in 300 people (1 in 222 in adults)[6] worldwide. According to the American Psychiatric Association (APA),

[6]https://www.who.int/news-room/fact-sheets/detail/schizophrenia, accessed 05/19/2023

| **Cohort/** Sex | Participants | Sessions | Mean Age; Std (Session Level) |
|---|---|---|---|
| **Controls** | | | |
| Female | 408 (63%) | 655 (62.8%) | 46.31; 16.37 |
| Male | 240 (37%) | 388 (37.2%) | 46.23; 16.03 |
| All | 648 | 1043 | 46.28; 16.24 |
| **Schizophrenia** | | | |
| Female | 10 (24.4%) | 19 (26.4%) | 36.11; 9.41 |
| Male | 31 (75.6%) | 53 (73.6%) | 36.58; 10.12 |
| All | 41 | 72 | 36.46; 9.87 |
| **Depression** | | | |
| Female | 66 (79.5%) | 76 (79.2%) | 34.61; 12.07 |
| Male | 17 (20.5%) | 20 (20.8%) | 35.00; 10.23 |
| All | 83 | 96 | 34.69; 11.66 |
| **BS ALS** | | | |
| Female | 38 (48.1%) | 67 (46.2%) | 61.72; 10.79 |
| Male | 41 (51.9%) | 78 (53.8%) | 61.28; 8.99 |
| All | 79 | 145 | 61.48; 9.83 |
| **BP ALS** | | | |
| Female | 31 (50%) | 54 (50.5%) | 58.07; 10.89 |
| Male | 31 (50%) | 53 (49.5%) | 62.15; 8.26 |
| All | 62 | 107 | 60.09; 9.85 |

Table 4: Cohort demographics.

active schizophrenia may be characterized by episodes in which the affected individual cannot distinguish between real and unreal experiences. The severity, duration and frequency of symptoms are highly variable, among patients as well as within the same individual over time (Buckley et al., 2008). Symptoms can be divided into three main categories: (1) positive (i.e. abnormally present) symptoms such as hallucinations or paranoia, (2) negative (i.e. abnormally absent) symptoms such as blunted affect (difficulty in expressing emotions) or anhedonia (inability to feel pleasure) and (3) disorganized symptoms which include abnormal movement as well as disordered thinking and speech.[7]

Among individuals with Schizophrenia, psychiatric and medical comorbidities such as substance abuse, anxiety and depression are common (Buckley et al., 2008; Green et al., 2003; Cassano et al., 1998). Buckley et al. (2008) point out that Depression is estimated to affect half of the patients. These comorbidities, as well as the variation in symptoms and medications, make the identification of true (multimodal) biomarkers for schizophrenia a difficult task.

As can be seen in Section 4.3, we assessed 41 individuals with a diagnosis of schizophrenia at

[7] https://www.psychiatry.org/patients-families/schizophrenia/what-is-schizophrenia, accessed 05/19/2023

a state psychiatric facility in New York, NY. The study was approved by the Nathan S. Kline Institute for Psychiatric Research and we obtained written informed consent from all participants at the time of screening after explaining details of the study. The assessment of both patients and controls was overseen by a psychiatrist.

### 4.3.2 Amyotrophic Lateral Sclerosis

ALS is a rare neurological disease that affects nerve cells in the brain and spinal cord that control voluntary muscle movement. The disease is progressive and there is currently no cure or effective treatment to reverse its progression.[8]. Global estimates of ALS range from 1.9 per 100,000 to 6 per 100,000.[9] In ALS patients, studies found comorbidity with dementia, parkinsonism and depressive symptoms (Körner et al., 2012). Diekmann et al. (2020) found depression to occur statistically significantly more often in ALS patients compared to Controls. In addition, Heidari et al. (2021) found in a meta-analysis of 46 eligible studies that the pooled prevalence of depression among individuals with ALS to be 34%, with mild, moderate, and severe depression rates at 29%, 16%, and 8%, respectively.

As shown in Section 4.3, data from 79 ALS bulbar symptomatic and 62 ALS bulbar pre-symptomatic patients were collected in cooperation with EverythingALS and the Peter Cohen Foundation[10]. In addition to the assessment of speech and facial behavior, participants filled out the ALS Functional Rating Scale-revised (ALSFRS-R), a standard instrument for monitoring the progression of ALS (Cedarbaum et al., 1999). The questionnaire comprises 12 questions about physical ability with each function's rating ranging from *normal function* (score 4) to *severe disability* (score 0). It includes four scales for different domains affected by the disorder: bulbar system, fine and gross motor skills, and respiratory function. The ALSFRS-R score is the total of the domain sub-scores, the sum ranging from 0 to 48. For this study, ALS patients were stratified into the following sub-cohorts based on their bulbar subscore: (a) BS ALS with a bulbar subscore $< 12$ (first three ALSFRS-R questions) and (b) BP ALS with a bulbar sub-score $= 12$.

### 4.3.3 Depression

Depression is a common mental health disorder characterized by persistent sadness and lack of interest or pleasure in previously enjoyable activities. In addition, fatigue and poor concentration

[8] https://www.ninds.nih.gov/health-information/disorders/amyotrophic-lateral-sclerosis-als, accessed 05/19/2023

[9] https://www.targetals.org/2022/11/22/epidemiology-of-als-incidence-prevalence-and-clusters/, accessed 05/19/2023

[10] https://www.everythingals.org/research

are common. The effects of depression can be long-lasting or recurrent and can drastically affect a person's ability to lead a fulfilling life. The disorder is one of the most common causes of disability in the world.[11] According to the APA, an estimated one in 15 adults (6.7%) is affected by depression each year. Moreover, one in six people (16.6%) will experience depression at some point in their lifetime.[12]
A well-established tool for assessing depression is the Patient Health Questionnaire (PHQ)-8 (Kroenke et al., 2009). Depression symptoms severity can be classified based on the following total score thresholds (Kroenke et al., 2001):

- 5 to 9: mild
- 10 to 14: moderate
- 15 to 19: moderately severe
- 20 to 24: severe

We investigated at least moderately severe depression cases, i.e. we classified depression cases based on a PHQ-8 of $>= 15$. The data for this study, including the completion of the PHQ-8, was collected through crowd-sourcing, resulting in a sample of 83 individuals that scored at or above this cutoff. Statistics for this cohort are summarized in Section 4.3.

[11] https://www.who.int/health-topics/depression, accessed 06/20/2023
[12] https://www.psychiatry.org/patients-families/depression/what-is-depression, accessed 06/20/2023

# 5 Methods

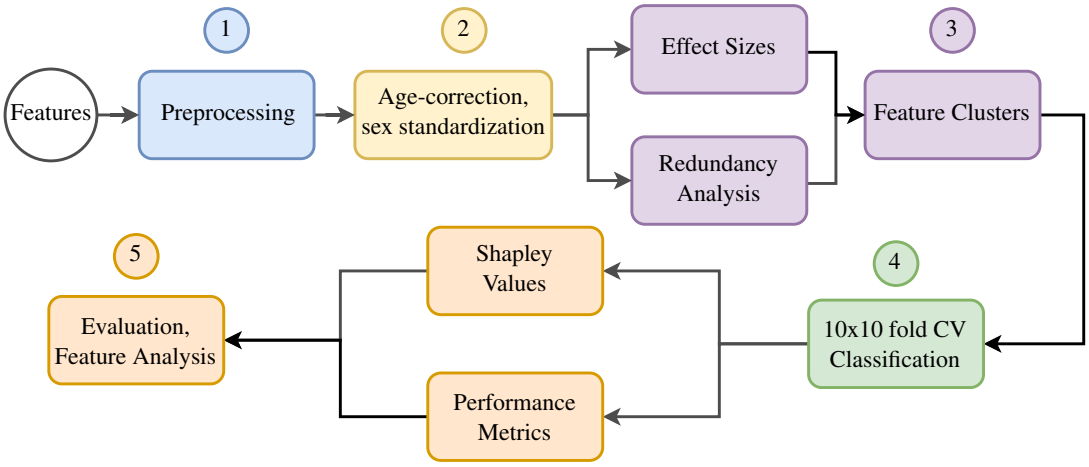



Figure 3: Overview of proposed feature selection pipeline.

Our approach aims at assessing disorder-relevant multimodal features by taking into account the factors of age and sex while minimizing redundancy, ensuring statistical meaningfulness and providing interpretability of discrimination experiments. Our procedure, which is shown in Figure 3, is divided into the following stages: (1) preprocessing, (2) age-correction and sex-normalization, (3) redundancy and effect size analysis, (4) classification, and (5) performance evaluation and feature analysis including the Shapley analysis.

## 5.1 Preprocessing

For a large part of the cohorts, we have information about their mental state assessed by the PHQ-8. To reduce the risk of confounding effects of depression, we filtered out all cases with a score of $>= 4$ from the control cohort. We did not have this information available for the ALS and Schizophrenia cohorts, including the control data collected in these studies. Furthermore, sessions with no information on sex and age were dropped. In addition, to acknowledge the differences between cohorts, we remove outliers for the healthy control dataset and the respective patient cohorts separately. Outliers may occur for a variety of technical, environmental or compliance reasons. For example, a loud barking dog in the background may skew the detected minimum or maximum frequency that the algorithm detects in the signal. Or, regarding facial features, a study participant chewing gum or moving their head out of frame will impact the accuracy of metrics or the detection of facial landmarks. To address such issues and remove the affected data points, we

apply an automatic and scalable method that is driven by the respective feature distributions. The outlier removal comprises the following steps:

1. Removal of feature values that are considered extreme outliers, meaning those that differ more than five standard deviations from the average of their respective group.
2. Recalculation of mean.
3. Removal of any feature values that still deviate more than three standard deviations from the mean.
4. Steps 2 and 3 are repeated recursively until no more features are found that deviate by this threshold.

To be able to compare feature sets across disorders, we identified and selected the tasks that are shared across datasets. For each participant some feature values may be missing for a variety of reasons such as failure in task completion, technical problems or processing issues. Since the used classifiers are not suited to work with missing values, an appropriate balance between the imputation of missing values and discarding data was required. As we aim at introducing as little uncertainty and noise, while preserving as much data as possible, we first filter both on the session and feature level. First, on the session level, we discard participant sessions that have more than 15% missing values. On the metric-task level, we filter out those features that have zero standard deviation, indicating that there is no variability across subjects, and those with more than 10% missing values. After those removal procedures, we impute missing values with mean feature values for the respective cohort in train and test sets separately.

## 5.2 Age-Correction & Sex-Normalization

Similar to the approach in Falahati et al. (2016), we apply a linear correction algorithm to both patient and control data based on age-related changes in the healthy control cohort only. By calculating age trends and coefficients on healthy controls, we aim to obtain the most accurate estimate of strictly age-related changes without the confounding effects of disease-related influences. In detail, for each feature, we fit a linear regression model to age as the independent and the feature as the dependent variable, modeling the age-related changes as a linear deviation. This is done separately for males and females to obtain a sex-specific result. Then, the sex-specific regression coefficients are used to remove the age-factor in all cohorts. More specifically, the corrected values are obtained by discounting the product of coefficient and age from the feature value of each

participant.
To account for sex-related differences, we applied sex-specific z-scoring to normalize the features. Z-normalization is a methodology that allows for the comparison or compilation of observations of different cohorts (Guilloux et al., 2011). In addition, the normalization process ensures the comparability of features on different scales by centering the feature distributions around 0 with a standard deviation of 1. First, the dataset to analyze was divided into male and female subjects. Then, each feature was normalized within each sex group using z-scoring. More specifically, mean and standard deviation of each feature were calculated within each sex group. The sample mean was then subtracted from each feature value, and the result was divided by the sample standard deviation to obtain the z-scores. This is reflected in the equation below, where $z$ represents the standardized value of $x$, $\bar{x}$ is the mean of the sample and $S$ is the standard deviation of the sample:

$$z = \frac{x - \bar{x}}{S} \tag{4}$$

The normalized features for males and females were merged back together into a unified dataset. To investigate the differences between male and female participants, we evaluate which features show statistical significance between males and females and calculate effect sizes to assess the magnitude of difference. Here, the MWU test is used for evaluating statistical significance ($p < 0.05$), Glass's Delta (Hopkins and Glass, 1978) to calculate effect sizes.

## 5.3 Redundancy Analysis & Effect Sizes

To identify collinear features and thereby assess redundancy, we perform hierarchical clustering on the Spearman rank-order correlations using the age-corrected and sex-normalized larger healthy control dataset. We apply the clustering for speech and facial features separately and merge the clusters to one set for combined analyses. The clustering procedure is motivated by the approach in Ienco and Meo (2008). It is based on Ward's method (Ward, 1963) which aims at minimising within-cluster variance. We implemented it using the scikit-learn library [13]. A dendrogram was plotted to inspect the feature groups visually. In hierarchical clustering, a dendrogram is a graphical representation that shows how data points, features in our case, or clusters that group these features are joined based on their similarity or dissimilarity. It is a tree-like structure illustrating merging and splitting of clusters based on different similarity thresholds. It can be used to understand the hierarchical relationships and clustering patterns within the data.
Based on the visual evaluation, we select a distance threshold that returns the most appropriate

[13] https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance_multicollinear.html

clustering for our purpose and based on metric-task specific knowledge to avoid clusters being too broad or too narrow. As we aim at selecting one feature per cluster within classification folds, the number of clusters determines the maximum number of features that are fed into the classifier. For this reason, we base our choice on two major factors: (a) balance between speech and facial clusters as we target roughly an equal number to avoid predominance of one modality over the other, and (b) common knowledge of task and feature domains (e.g. timing versus voice quality features, jaw versus eye movement or read versus free speech).
The clusters are used later on in the feature selection process in which we only consider the feature with the highest effect size for each cluster, if any at all. Effect sizes are calculated only for features showing statistical significance using the MWU test, as described in Section 5.2. We calculate those between the entire healthy control dataset and the respective cohorts to assess magnitude and direction of statistically valid differences between cases with a disorder and controls.

## 5.4 Classification

In this step, we run our feature selection pipeline with several well-known ML classifiers: Logistic Regression (LR), Multilayer perceptron (MLP) and RF. All classifiers are implemented using the scikit-learn library. The MLP with one hidden layer. We experimented with adding more hidden layers, but found that the minimal configuration with only one layer was beneficial in terms of performance. The hidden layer size is determined dynamically using the average of the number of selected features and the number of classes in the dataset. With regard to hyperparameters, the model is trained with a maximum of 10,000 iterations to allow sufficient time for convergence during training. Model training is stopped when the loss or score is not improving by a defined tolerance threshold. Here, we used the default of $1e-4$. Additionally, the alpha parameter is set to 0.001, controlling the regularization strength to prevent overfitting. The *sgd* (stochastic gradient descent) solver is utilized for optimization during training. The batch size is set to *auto*, enabling the model to determine the appropriate batch size during training. We further use the rectified linear unit function as the activation function.
Overall, the feature selection approach is designed to work with high-dimensional feature sets and small patient sample sizes when a large amount of healthy control data is available. The approach consists of several components: a demographic-inclusive preprocessing step, filter-based analysis involving statistical and correlation analysis, a post hoc embedded feature selection based on feature contribution in ML classifiers and a knowledge-based analysis utilizing consultation of clinical literature. Due to limited data availability, it was not practical at this time to split the data further to test the final feature set. Hence, further studies may explore the usefulness of

features derived including the step of embedded feature selection. An overview of the classification procedure is shown in **??**. First, feature selection and classification performance is evaluated for models aiming at classifying cases versus controls as well as between each combination of cases with a disorder.

Ten-fold cross-validation is applied to maximize the utilization of data for both training and testing purposes while mitigating overfitting to the classifier. To avoid bias towards the majority group, we create datasets that consist of an equal number of healthy controls versus cases with a condition. For each individual participant, we consider, if available, the first two sessions as independent data points. Using two sessions per participant may violate the assumption of independence of the MWU test if two sessions are not independent, leading to an increased risk of type I errors (false positive results). However, we hypothesize intra-participant observations to be as similar as inter-participant observations within the same group since an individual's performance may vary due to a variety of factors, including daily fluctuations in mood, fatigue, motivation, or other transient effects such as common colds. By considering the first two sessions, we take into account individual performance variability while minimizing the impact of possible learning effects. In addition, this approach enables us to increase the sample size and enhance statistical power. The larger sample size may allow for more precise estimates of population performance and reduces the risk of type II errors (false negatives). This ensures a more robust evaluation and improves the generalizability. For the classification experiments, we split the data using scikit-learn's *StratifiedGroupKFold* to make sure that sessions of the same participant are either in the respective training or testing fold. In each fold, we impute missing values and standardize features by sex using z-scoring as described in Section 5.2. This is done separately for training and test set. Then, we calculate for each feature if it differs statistically significantly between cases and controls, using a MWU test. For those that were found to differ, we calculate effect sizes, measured as Glass's delta. Glass's delta is calculated by dividing the difference between the means of the cohort groups, in our case cases with a disorder versus controls, by the standard deviation of the control group, yielding a standardized measure of effect size. This approach allows for possible differences in variability between the groups. As explained in Section 5.3, to minimize redundancy, we loop over each cluster of collinear features, and select the feature with the highest effect size, respectively. The selected features are fed to the classifier. In each classification fold, we calculate Shapley values of the selected features which we average across folds. A feature is counted for having a contribution value of 0 in a fold in which it was not selected at all. This experiment is run 10 times to smooth out performance variations and obtain more representative results.

In a second step, we perform multi-class classification, running feature selection across all possible binary combinations and take the union of features as input to the classifier. We assess how the

classifier performs with regard to differential diagnosis by evaluating the performance in terms of F1 score. The analysis of Shapley values is used for assessing which features the model focuses on for identifying which disorder in which particular set up. A more detailed explanation and description of the assessed evaluation metrics is given is Section 5.5.

## 5.5 Evaluation & Feature Analysis

To address the black-box nature of ML models such as the MLP, we aim for transparency and interpretability by leveraging Shapley values. By using Shapley values, we can gain insight into the decision-making process of the MLP as well as improve the interpretability of the less opaque LR and RF models. In this way, we can evaluate the importance of input features and better understand which input features are in the focus of the decision process.
We evaluate the performance of our feature selection approach by evaluating the classification performance of the selected markers as well as the robustness and clinical validity of the method. In the binary control-versus-disorder analyses, we employ a range of assessment metrics, including sensitivity, specificity, AUC, and the F1 score. These metrics collectively provide a comprehensive evaluation of the classifier's accuracy, sensitivity, specificity, and overall performance in distinguishing between healthy controls and cases with a disorder based on the selected features. For the across disorder studies, we use the F1 score as the only assessment metric. Unlike the classification tasks involving a control group, there is no clear reference class that can be considered as a positive or negative outcome. Hence, it becomes challenging to calculate sensitivity, specificity, or other evaluation metrics that rely on explicitly defining positive or negative classes. Similarly, for the multi-class experiments involving a control group and several disorders, we only utilize the F1 score. As for the binary across disorder experiments, there is no clear reference for a positive or negative outcome. Since the F1 score provides a balanced measure of precision and recall, it allows us to evaluate the model's performance across all classes equally. Additionally, we use a confusion matrix to document insights into the multi class model's predictions for each class. The following is an overview of the evaluation metrics used:

- **Sensitivity**: True Positive Rate (TPR): portion of diseased individuals detected as such.
- **Specificity**: True Negative Rate (TNR): portion of healthy individuals identified as such.
- **Receiver Operating Characteristic (ROC) curve**: Typically used as a measure of clinical utility, ROC curves show the performance of a model in terms of how sensitivity and specificity vary at each possible threshold by visualizing the trade-off between TPR and False Positive Rate (FPR; $1 - Specificity$).

- **Area Under the Curve (AUC)**: Calculates the performance of the model taking into account all possible thresholds that are visualized in the ROC curve.
- **Balanced F-Score**: As AUC scores may be misleading for small and/or imbalanced datasets (Hanczar et al., 2010), we further assess the balanced F-score which considers precision (portion of accurate predictions) and recall (= specificity) equally. This metric value is high when both component metrics are high. It is also useful when there is no clear definition of the positive or negative class.

To ensure robustness of the classification results, we employ a ten-fold cross-validation method repeated ten times with different random states. Features selected for classification are independent of the classifier. First, we assess the most important features by focusing on clusters selected in at least 85% of folds. This narrows our analysis to the most robust feature clusters across experiments while acknowledging the possibility of outlier results. Second, we calculate Shapley values for the best performing model, ranking features based on their relative importance. This approach prioritizes the model with the highest predictive performance, avoiding bias from underperforming models and maximizing the efficient use of computational resources. Third, we incorporate effect sizes of features into the analysis, assigning ranks accordingly. The rank sum of Shapley values and effect size ranks is calculated to determine the overall feature ranking. Finally, based on the top five identified features, we investigate the clinical literature for supporting evidence. This step enhances the interpretation of the findings by aligning them with existing knowledge.
We compare the classification performance of the proposed feature selection and classification pipeline to three baselines, each serving as a key example for one of the three feature selection categories:

1. Filter: Selection based on statistical significant difference of features between cases and controls, measured using the MWU test.
2. Wrapper: Recursive feature elimination, using LR as the estimator.
3. Embedded: Random forest classifier on the entire feature set.

For both (1) & (2) baselines, we use LR classifiers.
All baselines are tracked using the same metrics as the proposed feature selection approach, and are similarly run multiple times with different random seeds to investigate robustness with respect to the selected features across different training and test partitions, as well as the magnitude and stability of the performance metrics.

# 6 Results

## 6.1 Demographics Analysis

### 6.1.1 Age trends

To facilitate the comparison of coefficients on different scales, we report age trends as standardized linear regression coefficients ($\beta_{\text{std}}$) obtained by multiplying the coefficient ($\beta$) with the standard deviation of age ($\sigma_{\text{age}}$), the predictor variable, and dividing by the standard deviation of the respective feature, the output ($\sigma_{\text{feature}}$) (Menard, 2004; Siegel and Wagner, 2022). The formula can be expressed as follows:

$$\beta_{\text{std}} = \frac{\beta \cdot \sigma_{\text{age}}}{\sigma_{\text{feature}}} \tag{5}$$

This approach allows for a more interpretable and meaningful comparison of the impact of age across different features. It ensures that the coefficients represent the change in the outcome, i.e. the feature value, associated with an increase in age of one standard deviation. The calculations are performed separately for males and females, accounting for potential sex differences in the age trends. The standardized coefficients can be interpreted as effect sizes, where the following effect size magnitude thresholds apply (Cohen, 1988):

- small: $0.1 - 0.29$
- medium: $0.3 - 0.49$
- large: $>= 0.5$

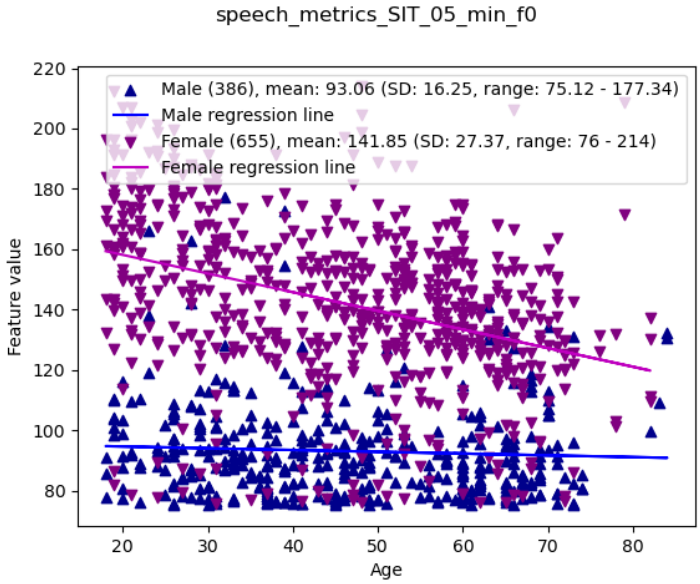
For clarity, we will use the term *effect size* consistently in this subsection, which we use synonymous for *standardized (linear) regression coefficient*.

| Domain | Metric | Effect size (standardized coefficient) & Tasks |
|---|---|---|
| Pitch | min f0 | -0.37 (SIT 05), -0.37 (SIT 07), -0.34 (SIT 15), -0.33 (SIT 11), -0.31 (SIT 09), -0.31 (SIT 13), -0.29 (Bamboo), -0.28 (Picture Description Task (PicDesc)) |
| Timing | articulation rate | -0.35 (Bamboo), -0.29 (SIT 11), -0.29 (SIT 15), -0.25 (SIT 07), -0.24 (SIT 13) |
| Timing | articulation time | 0.34 (Bamboo), 0.32 (SIT 11), 0.3 (SIT 15), 0.24 (SIT 07), 0.24 (SIT 13), 0.1 (SIT 05) |
| Timing | speaking rate | -0.34 (Bamboo), -0.32 (SIT 11), -0.28 (SIT 15), -0.27 (SIT 13), -0.19 (SIT 07) |
| Timing | speaking time | 0.32 (Bamboo), 0,32 (SIT 11), 0.27 (SIT 15), 0.26 (SIT 13) 0.21 (SIT 07) |
| Pitch | mean f0 | -0.3 (SIT 07), -0.29 (Bamboo), -0.26 (SIT 05), -0.25 (SIT 13), -0.25 (PicDesc), -0.24 (SIT 09), -0.24 (SIT 15), -0.22 (SIT 11) |
| Voice Quality | HNR | -0.21 (DDK), -0.19 (PicDesc), -0.16 (SIT 05), -0.16 (SIT 11), -0.16 (SIT 13), -0.13 (SIT 07), -0.11 (SIT 15), -0.1 (SIT 09) |
| Timing | CTA | 0.17 (SIT 11), 0.16 (Bamboo), -0.13 (SIT 05) |
| Pitch | stdev f0 | 0.19 (SIT 07), 0.17 (PicDesc), 0.16 (SIT 05), 0.15 (SIT 13), 0.15 (SIT 15), 0.1 (SIT 09) |
| Timing | PPT | 0.18 (SIT 11), 0.13 (SIT 13), 0.13 (SIT 15), 0.1 (Bamboo) |
| Voice quality | shimmer | 0.14 (SIT 05), 0.12 (PicDesc), 0.1 (SIT 07) |
| Energy | SNR | 0.12 (Bamboo), 0.11 (PicDesc) |
| Pitch | max F0 | -0.12 (SIT 05), -0.1 (SIT 11) |
| DDK-specific | syl.count | 0.11 |
| Voice quality | jitter | -0.11 (SIT 13) |

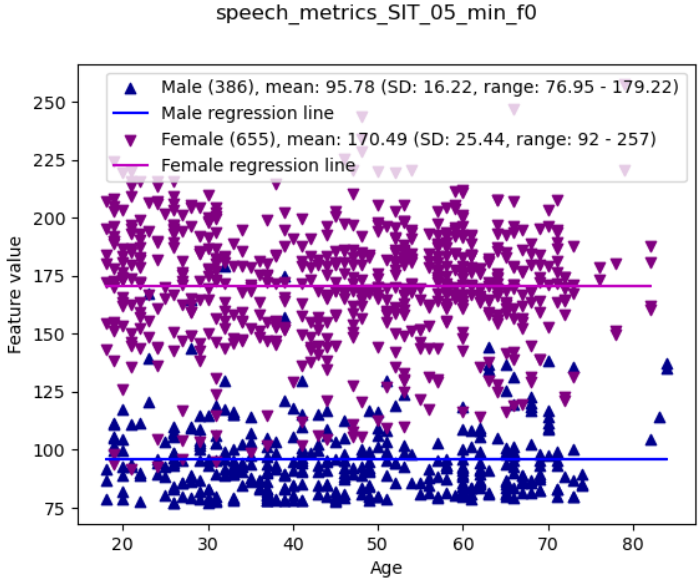
Table 5: Female age trends of speech features.

**Speech** Female age trends are shown in Table 5. Here, we observed the largest age-related trend in minimum F0 (e.g., -0.37 for SIT 5), as shown in Figure 4a, indicated by medium effect sizes across all SIT tasks and slightly below the medium threshold for the Bamboo reading passage (-0.29) and the picture description task (-0.28). In addition, we found a weaker negative trend for mean F0 (e.g., -0.3 for SIT 7) across these tasks. This suggests that the female voice becomes deeper with age.

The second largest age-related changes were found for the timing features articulation and speaking rate/time in Figure 5a. Our analysis shows that speech and articulation rate decrease with age, whereas speech and articulation time increase accordingly. In addition, PPT increases in the reading tasks involving longer passages (SIT 11, 13, 15, and Bamboo). Furthermore, we observed a lower HNR with age, the largest effect shown for the DDK task (-0.21) which suggests a change in voice quality. On the other hand, we observed higher SNR with increasing age, indicating a more prominent speech signal. In addition, we noted increased pitch variability with age, shown by an increasing standard deviation of F0. This implies that pitch variability becomes




(a) Age trends in min F0.




(b) Age-corrected feature values.

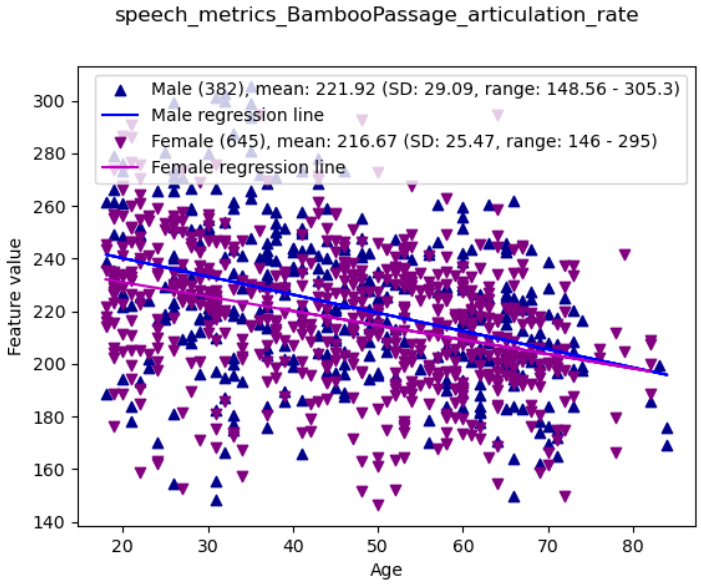
Figure 4: Speech feature with largest age-related trend in females before and after age-correction.

more pronounced with age. Some results, when effect sizes are small, should be taken with a grain of salt. For example, our analysis suggests a higher CTA with increasing age in some, but not all longer reading tasks and a negative trend for SIT 5 (-0.13).

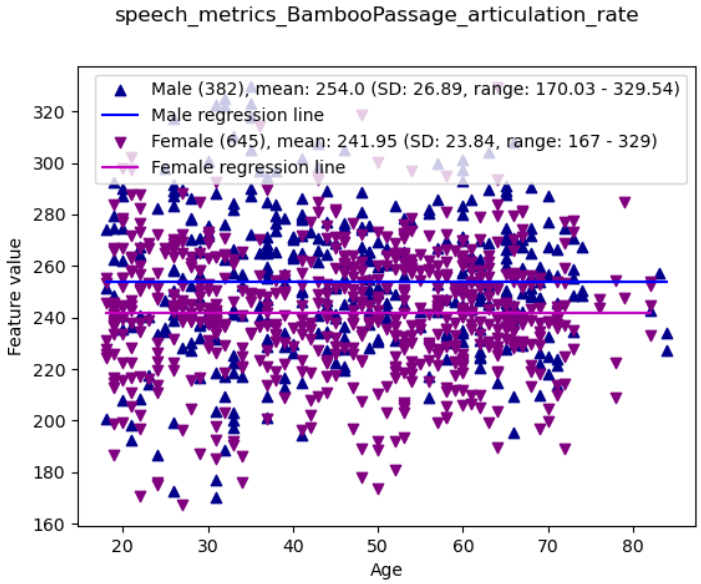
| Domain | Metric | Effect size (standardized coefficient) & Tasks |
|---|---|---|
| Timing | articulation time | 0.39 (Bamboo), 0.31 (SIT 15), 0.29 (SIT 11)<br>0.26 (SIT 07), 0.17 (SIT 13), 0.15 (PicDesc) |
| Timing | articulation rate | -0.38 (Bamboo), -0.31 (SIT 15), -0.26 (SIT 07), -0.25 (SIT 11),<br>-0.18 (SIT 13) |
| Timing | speaking time | 0.36 (Bamboo), 0.32 (SIT 11), 0.31 (SIT 15), 0.15 (PicDesc)<br>0.25 (SIT 07)), 0.21 (SIT 13), 0.1 (SIT 09) |
| Timing | speaking rate | -0.35 (Bamboo), -0.29 (SIT 11), -0.28 (SIT 15), -0.24 (SIT 07),<br>-0.19 (SIT 13) |
| Timing | CTA | 0.22 (Bamboo), 0.17 (SIT 13), -0.11 (SIT 05), 0.1 (SIT 11) |
| Energy | SNR | 0.21 (DDK), 0.15 (Bamboo), 0.14 (SIT 13), 0.1 (SIT 15) |
| DDK-specific | cTV | 0.18 |
| DDK-specific | syl. rate | -0.18 |
| Timing | PPT | 0.18 (SIT 11), 0.13 (Bamboo) |
| Pitch | stdev f0 | 0.18 (SIT 13), 0.17 (SIT 11), 0.16 (SIT 05), 0.15 (SIT 09),<br>0.14 (SIT 07), 0.11 (PicDesc) |
| Voice Quality | shimmer | 0.17 (SIT 05), 0.15 (DDK), 0.14 (PicDesc), 0.12 (SIT 09) |
| Pitch | max f0 | 0.16 (PicDesc), 0.15 (SIT 11), 0.15 (SIT 13), 0.14 (SIT 07),<br>0.12 (SIT 09) |
| Pitch | min f0 | -0.12 (SIT 07), -0.11 (SIT 09) |
| Voice Quality | jitter | -0.11 (SIT 11) |

Table 6: Male age trends of speech features.

For males, the largest age trends were found in timing metrics as shown in Table 6. With increasing age, participants exhibited longer articulation and speaking times, in particular across the longer reading passages such as the Bamboo reading task (e.g. 0.39 for articulation time, see Figure 5a). In line with that, we found that articulation and speaking rates decrease with age. The magnitude of age effects for these metrics are comparable to the ones observed in female participants.
In addition, older individuals exhibited a higher signal-to-noise ratio (SNR) across several tasks. Regarding DDK abilities, we discovered a decrease in syllable rate (-0.18) and an increase in cTV (0.18), which points to increasing difficulties in speech motor function with aging. The analysis of voice quality revealed that males exhibit higher shimmer levels with increasing age, such as in the picture description task (0.14), indicating greater variability in voice stability. In addition, age was associated with F0-related changes, measured in particular as a higher F0 standard deviation (e.g. 0.18 for SIT 13) and higher maximum F0 (e.g. 0.16 for the picture description task). This demonstrates a contrasting, albeit less pronounced, age-related effect in comparison to F0 related




(a) Age trends in articulation rate.




(b) Age-corrected feature values.

Figure 5: Speech feature with largest age-related trend in males before and after age-correction.

changes observed in female voices.
Figure 4b and Figure 5b show the distribution of features after the correction by age.

**Facial** Due to the large amount of facial features meeting the $>= 0.1$ threshold of small effect sizes, we only report facial features that show at least a medium age-related trend.

| Domain | Metric | Standardized Coefficient (ES) & Tasks |
|---|---|---|
| Mouth | avg. opening | -0.4 (Bamboo), -0.4 (SIT 11), -0.4 (SIT 13), -0.39 (DDK), -0.39 (SIT 05), -0.39 (SIT 07), -0.39 (SIT 09), -0.38 (SIT 15), -0.37 (PicDesc) |
| Eye | avg. opening | -0.36 (PicDesc), -0.32 (Bamboo), -0.32 (SIT 11), -0.32 (SIT 13), -0.31 (SIT 05), -0.3 (SIT 15) |
| Mouth | avg. symmetry | -0.33 (Bamboo), -0.31 (SIT 07), -0.31 (SIT 11), -0.3 (DDK), -0.3 (SIT 09), -0.3 (SIT 13), -0.3 (SIT 15) |
| Eye | max. opening | -0.3 (PicDesc), -0.3 (SIT 05) |

Table 7: Female age trends of facial features.

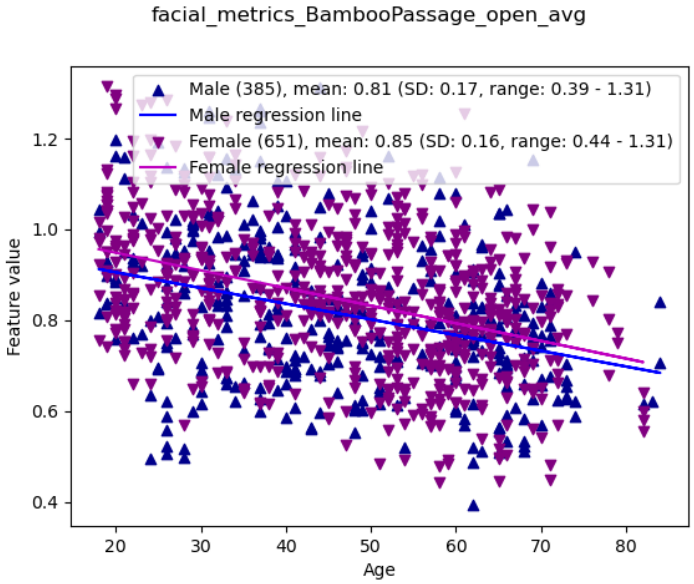
For female controls, we identified 257 facial features with an effect size $>= 0.1$, indicating an at least weak age-related effect. Speech features with at least a medium effect are shown in Table 7.

We identified the largest age-related trend as the lower average mouth opening in females across all tasks (e.g. -0.4 for the Bamboo task, as shown in Figure 6a). Similarly pronounced is the decline in average eye opening with increasing age, such as -0.36 in the picture description task. In addition, we find a decline in average mouth symmetry with aging, shown as negative effect sizes of medium magnitude across most tasks, where the largest effect is shown for the Bamboo reading passage (-0.33).

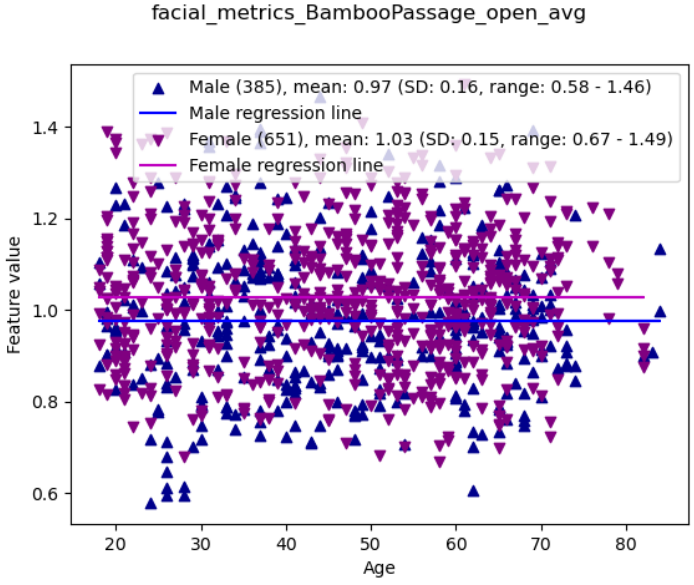
| Domain | Metric | Standardized Coefficient (ES) & Tasks |
|---|---|---|
| Mouth | avg. opening | -0.34 (DDK), -0.34 (SIT 11), -0.33 (Bamboo), -0.33 (SIT 09), -0.33 (SIT 13), -0.33 (SIT 15), -0.31 (SIT 07) |
| Jaw movement | abs. avg. jerk LL | 0.33 (SIT 13), 0.32 (SIT 09) |
| Jaw movement | abs. avg. speed LL | 0.32 (SIT 09), 0.32 (SIT 13), 0.31 (SIT 15) |
| Jaw movement | abs. avg. acc. LL | 0.32 (SIT 13), 0.31 (SIT 09), 0.31 (SIT 15) |
| Jaw movement | abs. max. jerk LL | 0.31 (SIT 13), 0.3 (Bamboo) |
| Jaw movement | min. jerk LL | -0.31 (SIT 13) |
| Jaw movement | abs. max. speed LL | 0.31 (SIT 13) |
| Jaw movement | abs. max. acc. LL | 0.3 (SIT 13) |
| Jaw movement | min. acc. LL | -0.3 (SIT 13) |
| Jaw movement | max. speed LL | 0.3 (SIT 13) |

Table 8: Male age trends of facial features.

We found 265 facial features with an effect size $>= 0.1$ for males. As for females, due to the high number of features, we report only male age-related trends with at least a medium effect, as shown in Table 8. We find a similar age-related trends in males compared to females for average mouth




(a) Age trends in articulation rate.




(b) Age-corrected feature values.

Figure 6: Facial feature with largest age-related trend in females before and after age-correction.

opening with the largest effect in the DDK task (-0.34). However, all other >= medium effects are shown for features concerning jaw movement, or more precisely, lower lip dynamics, which is not shown in this magnitude for females. Here, we find the largest age trends for jerk, speed and acceleration, particularly in the longer SIT tasks (e.g. 0.33 in SIT 13 for abs. avg. jerk LL).

### 6.1.2 Sex differences

For clarity, we have chosen to highlight the effect sizes of a selected subset of features. Note that these characteristics represent a limited number within each feature domain that serve as representative examples.

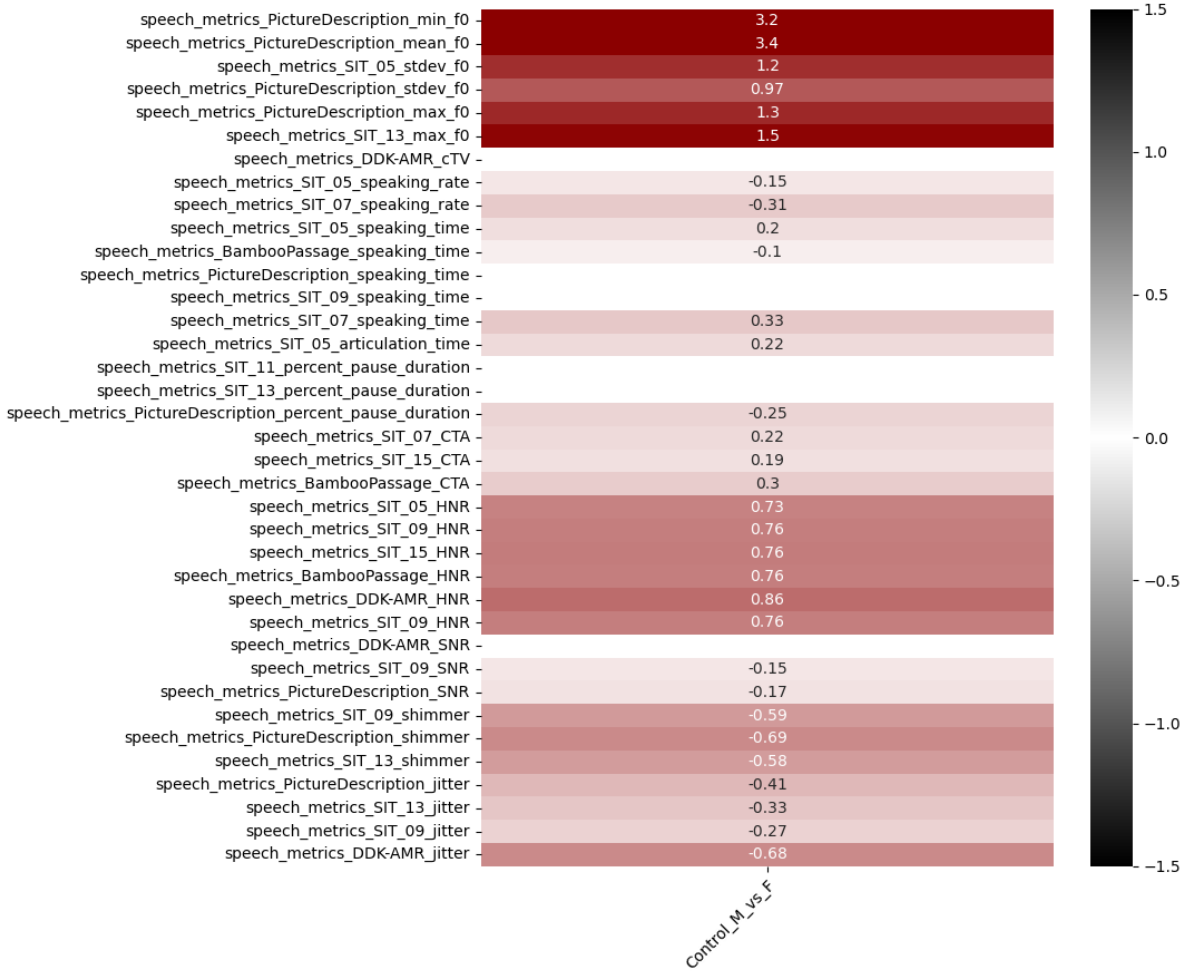



Figure 7: Effect sizes of statistically significantly different speech features between male and female controls. Positive effects indicate a larger value for females compared to males.

**Speech** As shown in Figure 7, differences between males and females are most pronounced in the frequency domain, concerning mean, maximum, minimum and standard deviation of the fundamental frequency.

For all frequency features investigated, effect size magnitudes between males and females are large. We find the largest effect for the mean F0 in the picture description task. Medium to large

effects are found for HNR and shimmer with the largest difference for HNR in the DDK task (0.86). In addition, we find lower jitter in female voices and higher CTA compared to males. For females, slower speaking is indicated by small effects regarding a lower speaking rate and higher speaking time for some tasks. We also find a statistically significantly lower PPT for females in the picture description task.
Overall, we observe that the majority of speech features investigated shows significant differences between males and females.

**Facial** For facial features, we identify overall few differences and small effect size magnitudes between males and females. We found a statistically significantly higher mean (0.26) and max (0.25) mouth surface area for the SIT 7 and a wider lip aperture in the picture description task (0.31) in females compared to males. All other identified statistically significant differences fall below the threshold of a small effect size.

## 6.2 Redundancy Analysis

The clustering analysis was performed by first generating a dendrogram, regardless of a distance threshold. Then, a threshold is set to divide the clusters. These distance thresholds have no unit of measurement or specific scale as prior to performing clustering using Ward's linkage, the correlations between features are converted into distances by subtracting the absolute value of the correlation from 1, yielding a measure where larger correlations result in smaller distances between features. As a consequence, distance values represent distances relative to other clusters, with smaller values indicating higher correlations. Since our speech and facial dendrograms include a large number of features, it is not feasible to display them in their entirety in this thesis. A selected section of the speech dendrogram is shown in Figure 8. In this figure, the light gray horizontal lines represent 0.1-distance increments with the red line signaling a threshold of 2. The green and blue lines show the connections between features and formed clusters. By observing the distance thresholds at the corresponding cluster links, the specific points at which clusters merge or split can be identified relative to each other.
We present the identified clusters for speech and facial features separately. We targeted an approximately equal number of clusters.

**Speech** Regarding the grouping of speech features, we manually adjusted the distance threshold to 1.1 after visual inspection of the dendrogram and potential clusters based on different thresholds. This yielded a total of 13 distinct speech clusters as can be seen in Table 9. Among these clusters,

| # | Cluster domain | Metrics | Tasks | # Features |
|---|---|---|---|---|
| 1 | Energy | SNR | all | 8 |
| 2 | Timing alignment | CTA | all | 6 |
| 3 | Timing, pauses | PPT | all | 5 |
| 4 | Timing, speaking (1) | articulation/speaking time | Picture Description | 2 |
| 5 | Energy & articulation skills | SNR, syl.rate, syl.count & cTV | DDK | 4 |
| 6 | Timing, speaking (2) | articulation/speaking rate/time | SIT 5 & 9 | 8 |
| 7 | Timing, speaking (3) | articulation/speaking rate/time | SIT 7, 11, 13, 15, Bamboo task | 21 |
| 8 | Voice quality (DDK skills) | HNR, jitter & shimmer | DDK | 3 |
| 9 | Voice quality (periodicity) | HNR | all except DDK | 8 |
| 10 | Voice quality (amplitude variation) | shimmer | all except DDK | 8 |
| 11 | Voice quality (frequency variation) | jitter | all except DDK | 8 |
| 12 | Frequency (mean, min) | min & mean F0 | all | 16 |
| 13 | Frequency (max, std) | max & std F0 | all | 16 |

Table 9: Speech clusters identified by hierarchical clustering.

five were associated with timing, four with voice quality, two with frequency, one DDK-specific (energy and articulation skills), and one with energy-related features. One cluster specifically comprised features relevant to the DDK task, such as syllable rate, count, and cTV.

Overall, the DDK task metrics tended to form separate clusters, while the reading tasks of SIT and Bamboo, which are closely related, were grouped together most often in multiple feature domains. At feature domain level, the redundancy analysis identified the features of the *voice quality* domain as being the highest correlated to each other compared to other domains. As can be seen in Figure 8, all voice quality metrics pertaining to the DDK task are identified as a cluster below a threshold of 1.2, while for all other tasks the metrics of HNR, jitter, and shimmer each form a separate cluster. To account for the different nature of diadochokinetic features compared to reading or free speech features, this was key to the choice of the 1.1 threshold, in addition to the number of clusters and the meaningfulness of the other categories.

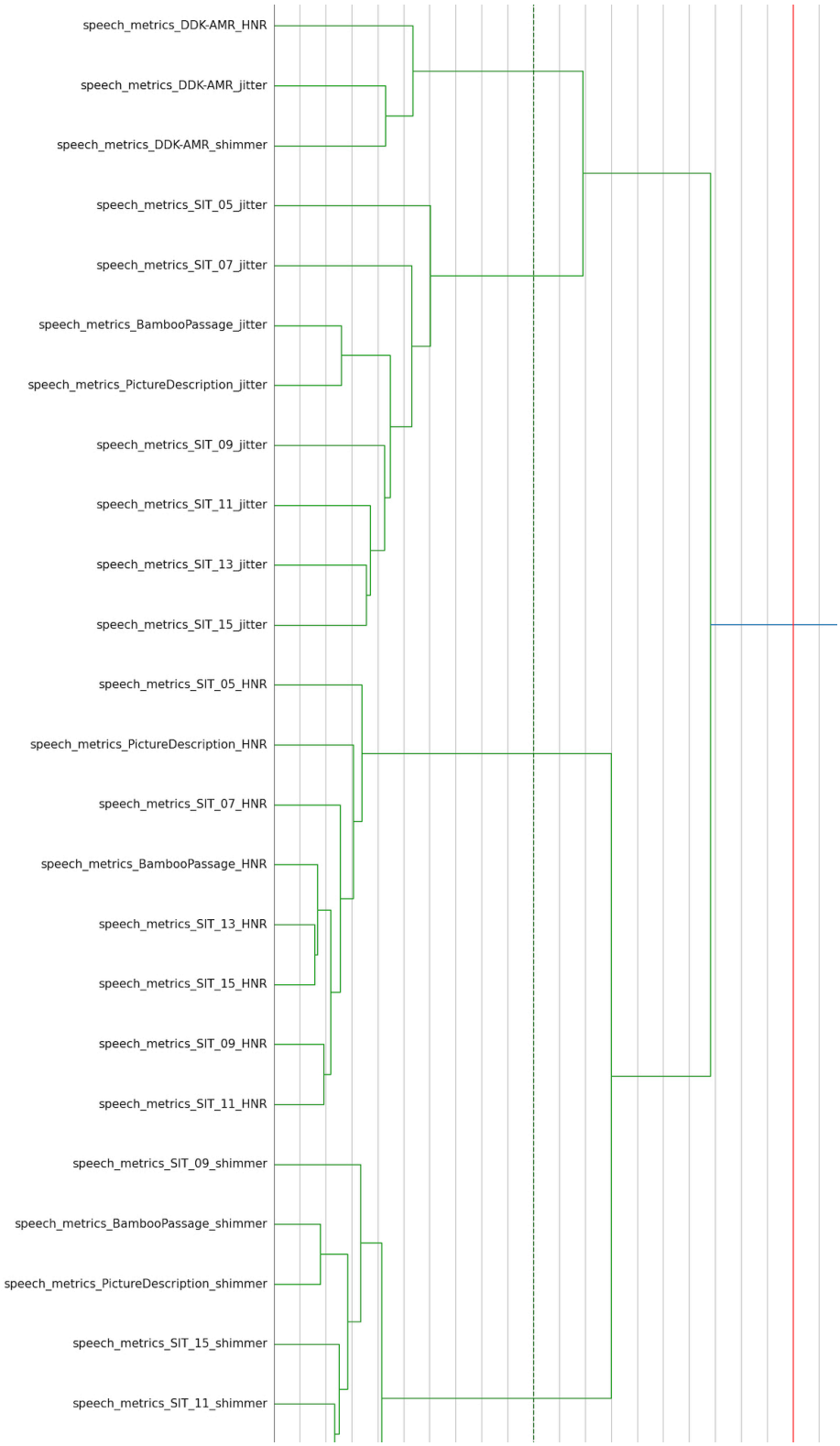



Figure 8: Dendrogram cutout showing voice quality clustering.

In some cases, multiple feature clusters related to the same domain such as timing (speaking and articulation features) which is reasonable considering the different nature in tasks. On the feature level overall, speaking and articulation timing features are found to differ the most based on the specific task.

| # | Cluster domain | Metrics | Tasks | # Features |
|---|---|---|---|---|
| 1 | Lip movement (1) | speed, acc. & jerk measures | all except DDK | 95 |
| 2 | Lip width | mean & max lip width | all | 18 |
| 3 | Mouth opening | mean & max lip aperture, mouth surface area | all | 36 |
| 4 | Lip movement (2) | speed, acc. & jerk metrics | DDK | 12 |
| 5 | Jaw movement (1) | speed, acc. & jerk metrics | DDK | 12 |
| 6 | Jaw movement (2) | speed, acc. & jerk metrics | SIT 7 | 12 |
| 7 | Jaw movement (3) | speed, acc. & jerk metrics | SIT 5 | 12 |
| 8 | Jaw movement (4) | min + max speed, acc. & jerk metrics | Picture Description | 9 |
| 9 | Jaw movement (5) | speed, acc. & jerk metrics | SIT 9, 11, 13, 15, Bamboo, Picture Description (mean) | 63 |
| 10 | Mouth symmetry | mean mouth symmetry | all | 9 |
| 11 | Eye opening | mean and max eye opening | all | 18 |

Table 10: Facial clusters identified by hierarchical clustering.

**Facial** The visual inspection of the dendrogram suggested a distance threshold of 1.7 and resulted in a total of 11 facial clusters as shown in Table 10. Identified were the following clusters: two relating to lip movement features, one for lip width, one mouth opening, five concerning jaw dynamics, one mouth symmetry and one cluster comprising all eye opening features. Notably, as in the speech domain, the DDK task exhibited a higher tendency to form distinct clusters compared to other tasks, with lip and jaw movement features clustered separately for DDK. On the domain level, the clustering identified most differences based on tasks within jaw movement features. Here, DDK and the two shortest sentence intelligibility tests formed their own clusters, respectively. Furthermore, the longer speaking tasks, reading (SIT & Bamboo as well as the average metrics of the Picture Description task) were grouped together. Furthermore, within the picture description task, measures of extreme movement (minimum or maximum) formed a distinct cluster. In terms of metrics nature (dynamics versus surface measures), we found that those were consequently separated into different groups. Additionally, the algorithm maintained separate clusters for different domains (e.g., jaw versus eyes) and only identified domain-specific subgroups.

## 6.3 Effect Sizes

Positive effect sizes represent that feature values are larger for cases with a disorder than controls. Conversely, negative values represent larger feature values for controls than cases with a disorder. Commonly used effect size magnitude thresholds as suggested in Cohen (1988) are:

- small: $0.2 - 0.5$
- medium: $0.5 - 0.8$
- large: $> 0.8$

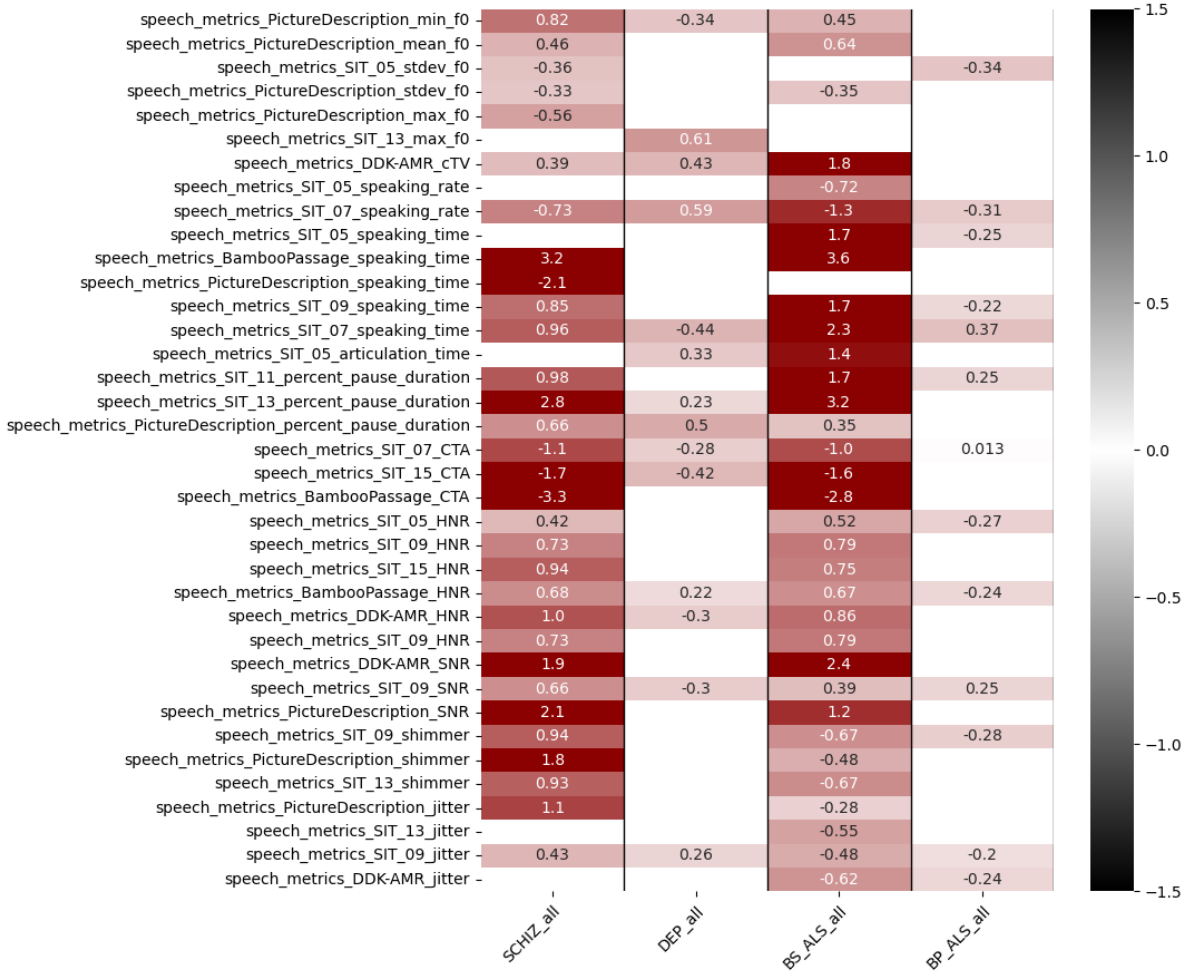



Figure 9: Effect sizes of age-corrected and sex-normalized speech features between cases with a disorder and controls.

**Speech** We observe speech features with large effect sizes between cases with a disorder and controls for BS ALS and schizophrenia cohorts and overall smaller effect sizes for depression and BP ALS cases, as shown in Figure 9.

The largest effects in schizophrenia are shown for CTA (-3.3 for the Bamboo reading passage), speaking time (3.2 for the Bamboo reading passage and -2.1 for the Picture Description task), PPT (2.8 for SIT 13) and SNR (2.1 for Picture Description). This shows that patients exhibit a lower CTA, referring to the synchronization between their own speech and the one of the virtual agent. In addition, they pause longer when reading sentences and speak louder than controls. Notably, they need more time to complete the Bamboo reading task, while exhibiting a slower speaking rate, but speak shorter in the free speech task, indicating a potential speech impairment or reluctance.

The largest effect in depression cases is shown by a higher maximum F0 (0.62 for SIT 13) compared to controls, higher speaking rate (0.59 for SIT 7) combined with a shorter speaking time and a higher PPT in the Picture Description task.

BS ALS patients exhibit a higher speaking time (3.6 for the Bamboo task) and longer pauses (3.2 for SIT 13) in reading tasks. They also show a lower synchronization with the virtual agent's reading patterns indicated by the CTA (-2.8 for the Bamboo task) and a louder voice in the DDK task (2.4).

BP ALS patients show overall much smaller effects than BS ALS patients. The largest effects which, however, all represent small effect sizes, are revealed in a longer speaking time and slower speaking rate (0.37 and -0.31 for the SIT 7) as well as a lower F0 standard deviation in the SIT 5 reading task (-0.34), indicating less pitch variation.

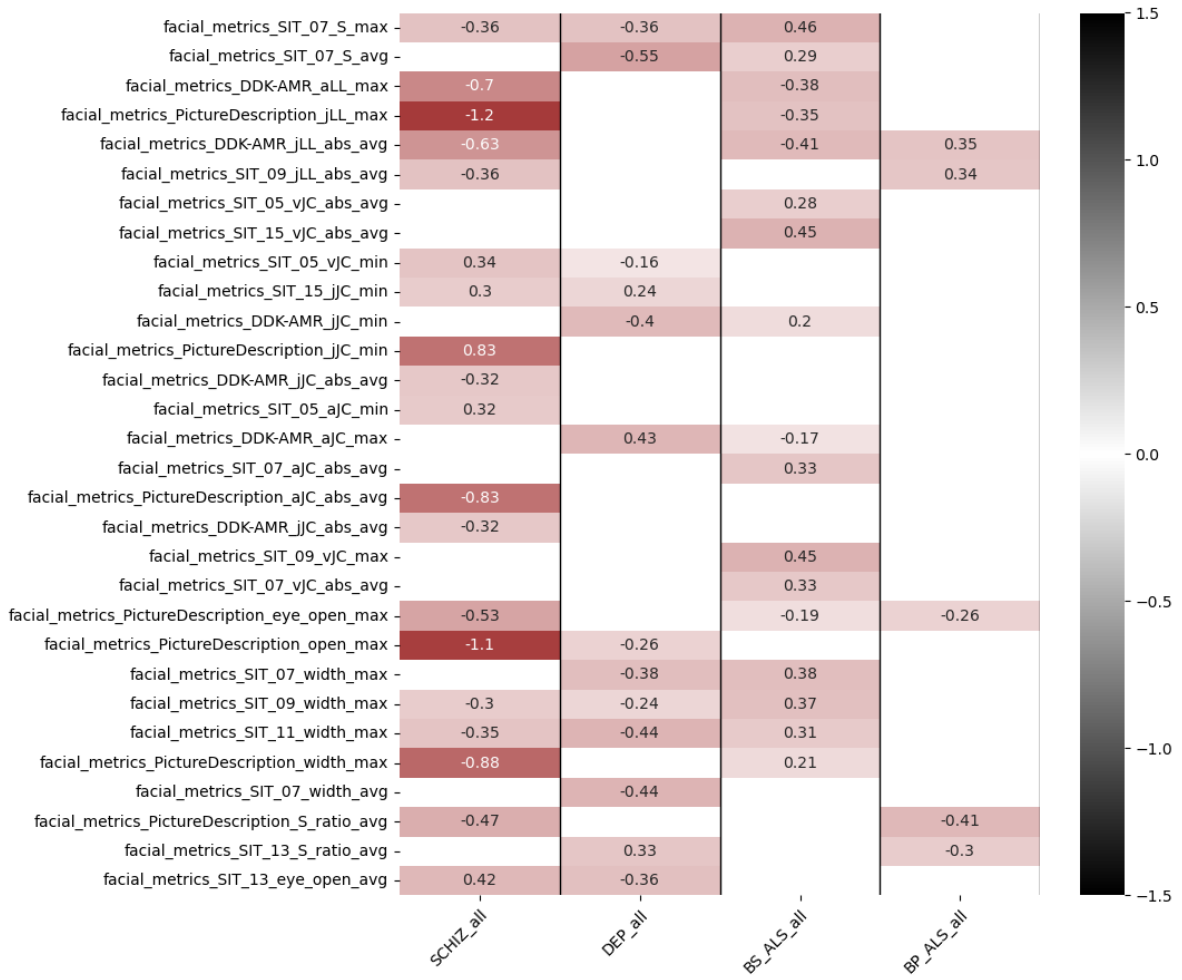



Figure 10: Effect sizes of age-corrected and sex-normalized facial features between cases with a disorder and controls.

**Facial** We found overall lower effect size magnitudes for facial features compared to speech features, as can be seen in Figure 10. The largest differences in facial features between cases with a disorder and controls are observed for the schizophrenia cohort. Large effects are revealed among lower lip (e.g. -1.2 for LL jerk in the picture description task) and jaw movement, lower maximum eye and mouth and smaller maximum lip width features. In all other cohorts, we find a maximum of medium effects. Fewest differences compared to controls are revealed for BP ALS cases.

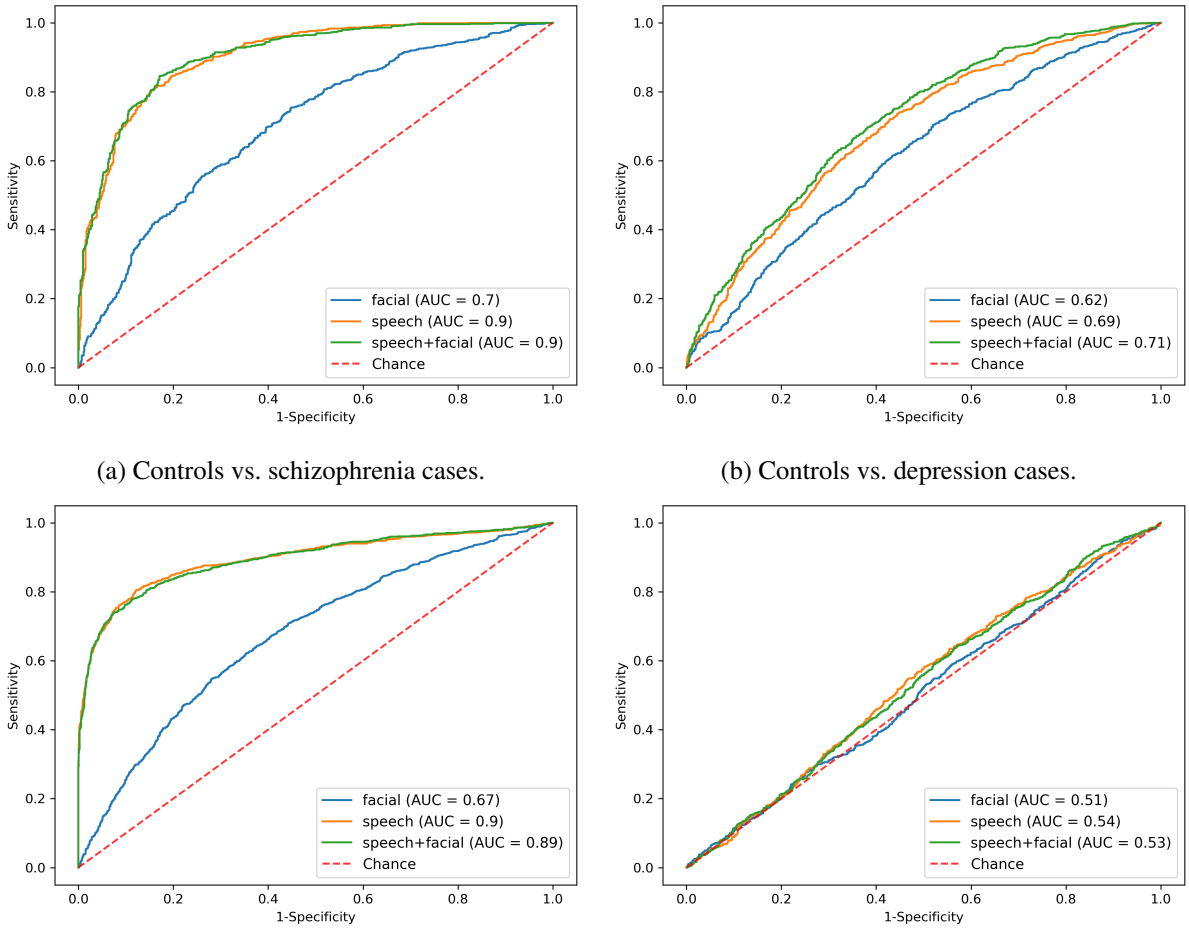
## 6.4 Classification & Shapley Values

### 6.4.1 Binary Classification

| Cohort | Evaluation | FS + LR | FS + RF | FS + MLP | Filter BL | Wrapper BL | Embedded BL |
|---|---|---|---|---|---|---|---|
| Depression | Sensitivity | **0.65** | **0.65** | **0.65** | 0.64 | 0.6 | 0.62 |
| | Specificity | 0.65 | 0.66 | 0.65 | 0.64 | 0.6 | **0.69** |
| | AUC | 0.71 | 0.71 | 0.71 | 0.69 | 0.64 | **0.72** |
| | F1 | **0.65** | **0.65** | **0.65** | 0.64 | 0.6 | **0.65** |
| Schizophrenia | Sensitivity | 0.83 | **0.85** | **0.85** | 0.84 | 0.84 | 0.83 |
| | Specificity | 0.8 | 0.77 | 0.82 | 0.83 | **0.85** | 0.79 |
| | AUC | 0.9 | 0.89 | 0.9 | 0.9 | **0.92** | 0.88 |
| | F1 | 0.82 | 0.81 | 0.83 | 0.83 | **0.85** | 0.81 |
| BS ALS | Sensitivity | 0.82 | **0.84** | 0.82 | 0.82 | 0.79 | 0.8 |
| | Specificity | 0.82 | 0.8 | 0.83 | 0.78 | **0.84** | 0.82 |
| | AUC | 0.89 | **0.9** | 0.89 | 0.88 | 0.88 | 0.89 |
| | F1 | 0.82 | 0.82 | **0.83** | 0.8 | 0.81 | 0.81 |
| BP ALS | Sensitivity | 0.52 | 0.51 | 0.52 | 0.52 | 0.52 | **0.55** |
| | Specificity | 0.54 | 0.54 | 0.53 | 0.54 | **0.58** | 0.56 |
| | AUC | 0.54 | 0.53 | 0.53 | 0.54 | 0.56 | **0.58** |
| | F1 | 0.53 | 0.52 | 0.52 | 0.53 | **0.55** | **0.55** |

Table 11: Binary classification results for baselines and the feature selection pipeline (cases with a disorder versus controls). In each row, we highlighted the highest performance in the particular evaluation metric. FS: Feature selection pipeline, BL: Baseline

Performances between classifiers implementing our feature selection pipeline and baselines are very similar (see Table 11). Regarding the implementation of the custom feature selection pipeline, there is no classifier that outperforms all others across disorders. However, the filter-based baseline model is outperformed by all other approaches. Among the wrapper and embedded feature selection baseline models, embedded feature selection worked best for depression and BP ALS classification while for schizophrenia, wrapper feature selection performed best.
For the sake of clarity and simplicity, we focus on the MLP results for the following in-depth analysis of performance across modalities and for feature analysis.




(a) Controls vs. schizophrenia cases.

(b) Controls vs. depression cases.

(c) Controls vs. BS ALS cases.

(d) Controls vs. BP ALS cases.

Figure 11: ROCs: Controls versus cases with a disorder.

ROC curves for control versus cases with a disorder classification experiments are shown in Figure 11. Overall, the classification performance between controls and schizophrenic as well as BS ALS cases is best. For distinguishing depression from healthy controls, the performance is lower, though also well above chance. For BP ALS, the results indicate that the classifiers, independent of the feature selection method, are not able to generalize based on the information provided. In depression, combining speech and facial modalities resulted in improved classification performance compared to speech or facial features alone, as shown in Figure 11. However, adding facial information did not enhance performance for schizophrenia and ALS cohorts compared to utilizing speech features alone.

We conduct a comprehensive analysis to determine the significance of features for distinguishing between cases with a disorder and controls using three key criteria. First, we assess the frequency

with which feature categories are selected across classification folds, restricting the set for further analysis to those that occur in $>= 85\%$ of folds. Second, we rank their contribution to the model output based on Shapley values obtained from the best performing model and effect sizes. By assigning ranks within each criterion, we calculate the rank sum, whereby a lower sum indicates a more crucial feature. To provide a more in-depth understanding, we discuss the most important features for each disorder in Section 7.

| | Speech | Facial | Speech & Facial |
|---|---|---|---|
| Sensitivity | 0.64 | 0.58 | **0.65** |
| Specificity | 0.64 | 0.59 | **0.65** |
| AUC | 0.69 | 0.62 | **0.71** |
| F1 | 0.64 | 0.59 | **0.65** |

Table 12: Performance metrics: Depression cases versus controls.

**Controls vs. depression** The performance metrics shown in Figure 11b and Table 12 for depression case classification using features selected of the speech, facial, and combined modalities demonstrate promising results, with the combined modalities showing the highest performance across evaluation metrics. More specifically, the combination of speech and facial features achieved the highest sensitivity (0.65), specificity (0.65), AUC (0.71), and F1 score (0.65) compared to the single modality models. These findings suggest that utilizing both speech and facial modalities can improve the detection of depression cases when compared to controls.

| Cluster | Feature | SHAP rank | ES (rank) | Rank sum |
|---|---|---|---|---|
| **Timing, speaking (#3)** | **SIT_07_speaking_rate (S)** | 1 | 0.59 (1) | 2 |
| **Timing, pauses** | **PictureDescription_PPT (S)** | 3 | 0.5 (3) | 6 |
| **Energy & articulation skills** | **DDK-AMR_cTV (S)** | 2 | 0.43 (5) | 7 |
| **Lip width** | **PictureDescription_width_avg (F)** | 6 | -0.44 (4) | 10 |
| **Eye opening** | **SIT_13_eye_open_avg (F)** | 7 | -0.36 (5) | 12 |
| Mouth opening | SIT_07_S_avg (F) | 12 | -0.55 (2) | 19 |
| Timing alignment | SIT_13_CTA (S) | 15 | -0.31 (6) | 21 |

Table 13: Depression: Most important features.

Section 6.4.1 shows the most important features across experiments. As shown here, we identify features of the timing domain, more specifically a higher speaking rate (SIT 7) and higher PPT (PicDesc) as well as a higher cTV (DDK task) and two facial features concerning eye opening and lip width among the most important features. We observed that these features made significant

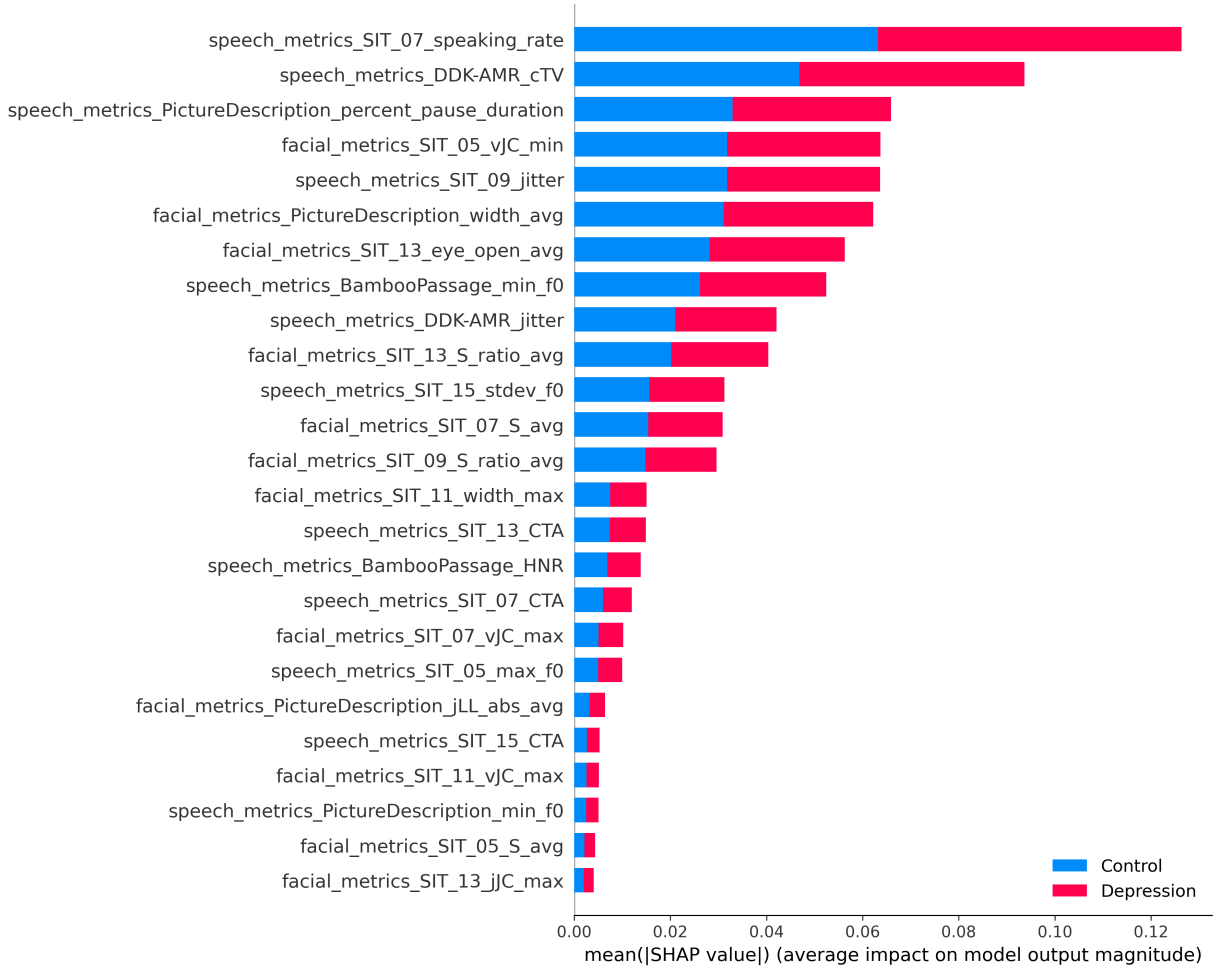



Figure 12: Shapley values: Controls versus depression cases.

contributions to the classifiers' predictions, as evidenced by their strong influence indicated by their Shapley values (see Figure 12), and also exhibited substantial effect sizes when comparing depression cases and controls, as revealed in Section 6.3.

| | Speech | Facial | Speech & Facial |
|---|---|---|---|
| Sensitivity | 0.84 | 0.65 | **0.85** |
| Specificity | 0.81 | 0.63 | **0.82** |
| AUC | **0.9** | 0.7 | **0.9** |
| F1 | 0.82 | 0.64 | **0.83** |

Table 14: Performance Metrics: Schizophrenia cases versus Controls.

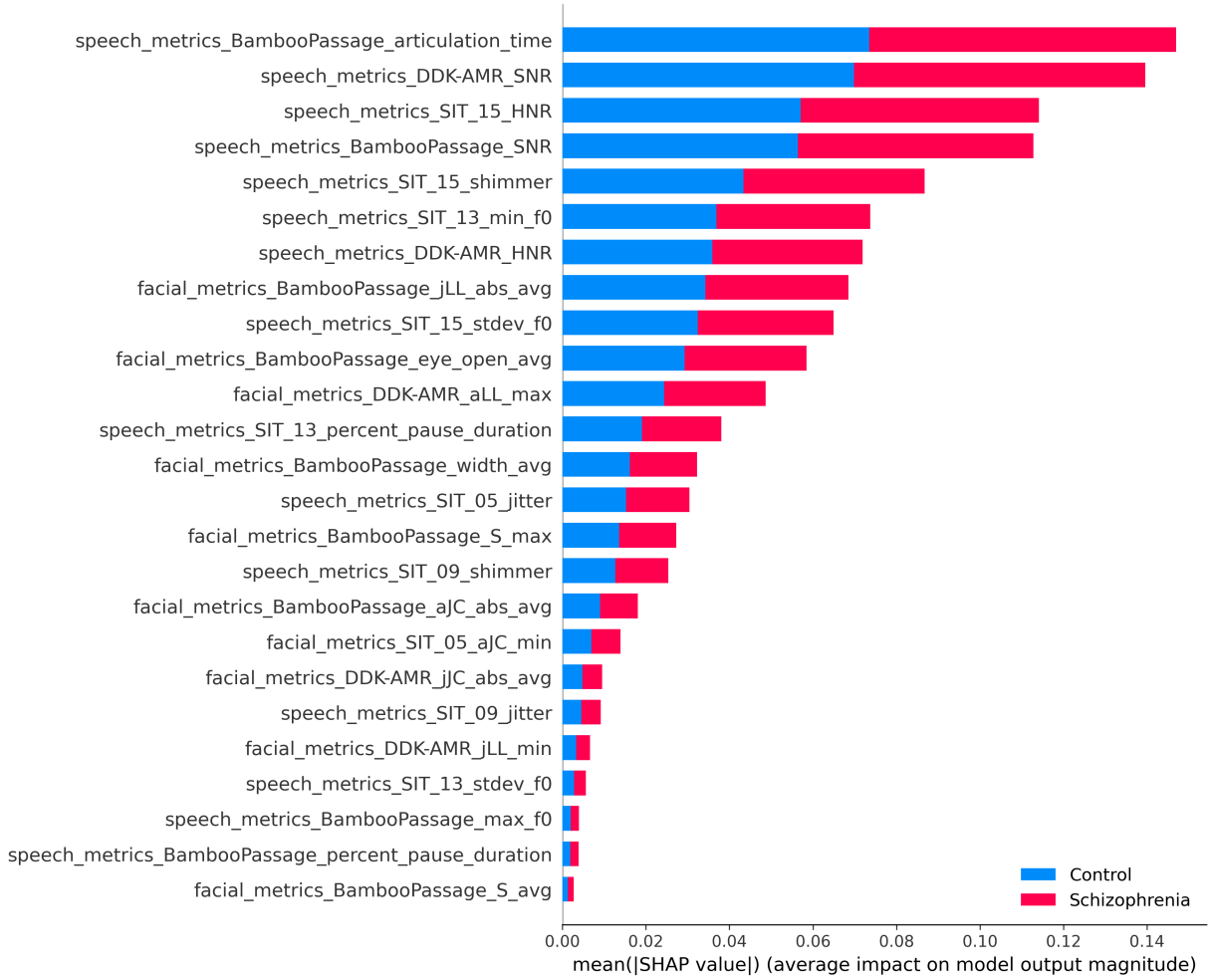



Figure 13: Shapley values: Controls versus Schizophrenia.

| Cluster | Feature | SHAP rank | ES (rank) | Rank sum |
|---|---|---|---|---|
| **Timing, speaking (3)** | **BambooPassage_articulation_time (S)** | 1 | 3.2 (1) | 2 |
| **Energy & articulation skills** | **DDK-AMR_SNR (S)** | 2 | 1.9 (3) | 5 |
| **Energy** | **BambooPassage_SNR (S)** | 4 | 1.0 (4) | 8 |
| **Voice quality (periodicity)** | **SIT_15_HNR (S)** | 3 | 0.94 (6) | 9 |
| **Voice quality (amplitude variation)** | **SIT_15_shimmer (S)** | 5 | 1.0 (4) | 9 |
| Voice quality (DDK skills) | DDK-AMR_HNR (S) | 7 | 1.0 (4) | 11 |
| Lip movement (1) | BambooPassage_jLL_abs_avg (F) | 8 | -0.96 (5) | 13 |
| Timing, pauses | SIT_13_percent_pause_duration (S) | 12 | 2.8 (2) | 14 |
| Frequency (max, std) | SIT_15_stdev_f0 (S) | 9 | -0.46 (9) | 18 |
| Lip movement (2) | DDK-AMR_aLL_max (F) | 11 | -0.7 (7) | 18 |
| Jaw movement | BambooPassage_aJC_abs_avg (F) | 17 | -0.56 (8) | 25 |
| Voice quality (frequency variation) | SIT_09_jitter (S) | 20 | 0.43 (10) | 30 |

Table 15: Schizophrenia: Most important features.

**Controls vs. schizophrenia** As can be seen in Table 14, speech features performed better than facial features for this task, and the combination of both modalities yielded marginal gains. Section 6.4.1 shows the features identified as most important in schizophrenia. In particular, we determined a high feature contribution, as represented by the Shapley values in Figure 13 and high effect sizes in articulation time as well as energy and voice quality features.

| | Speech | Facial | Speech & Facial |
|---|---|---|---|
| Sensitivity | 0.81 | 0.64 | **0.82** |
| Specificity | **0.87** | 0.62 | 0.83 |
| AUC | **0.9** | 0.67 | 0.89 |
| F1 | **0.84** | 0.63 | 0.83 |

Table 16: Performance Metrics: BS ALS cases versus Controls.

**Controls vs. BS ALS** As shown in Table 16, for speech, high values were observed for sensitivity, specificity, AUC and F1 scores, while specificity (0.87) was shown to be notably higher than sensitivity (0.81). This suggests that the classifier demonstrates a higher capability in correctly identifying healthy individuals who do not have the disorder, thus minimizing false positive results, compared to its ability to accurately detect individuals with the disease, i.e., predicting true positives. Facial features performed inferior to speech features in all evaluation metrics. Combining speech and facial features demonstrated slightly improved performance for sensitivity, while specificity, AUC and the F1 score showed a slightly lower performance compared to the speech modality alone.

| Cluster | Feature | SHAP rank | ES (rank) | Rank sum |
|---|---|---|---|---|
| **Timing, pauses** | **SIT_13_percent_pause_duration** | 3 | 3.2 (2) | 5 |
| **Timing alignment** | **BambooPassage_CTA** | 2 | -2.8 (3) | 5 |
| **Timing (3)** | **BambooPassage_speaking_time** | 4 | 3.6 (1) | 5 |
| **Energy & articulation skills** | **DDK-AMR_SNR** | 1 | 2.4 (4) | 5 |
| **Timing, speaking (2)** | **SIT_09_speaking_time** | 6 | 1.7 (5) | 11 |
| Voice quality (DDK skills) | DDK-AMR_HNR | 9 | 0.86 (7) | 15 |
| Lip movement (2) | DDK-AMR_jLL_abs_avg | 8 | -0.41 (11) | 19 |
| Energy | PictureDescription_SNR | 15 | 1.2 (6) | 21 |
| Frequency (mean, min) | PictureDescription_mean_f0 | 19 | 0.64 (9) | 28 |
| Frequency (max, std) | PictureDescription_stdev_f0 | 17 | -0.35 (12) | 29 |
| Voice quality (frequency variation) | SIT_09_jitter | 22 | -0.48 (10) | 32 |
| Voice quality (periodicity) | SIT_09_HNR | 25 | 0.79 (8) | 33 |

Table 17: BS ALS: Most important features.

The feature selection analysis shown in Section 6.4.1 suggests a predominant importance of timing categories, including pauses (PPT in SIT 13), timing alignment (CTA in the Bamboo reading

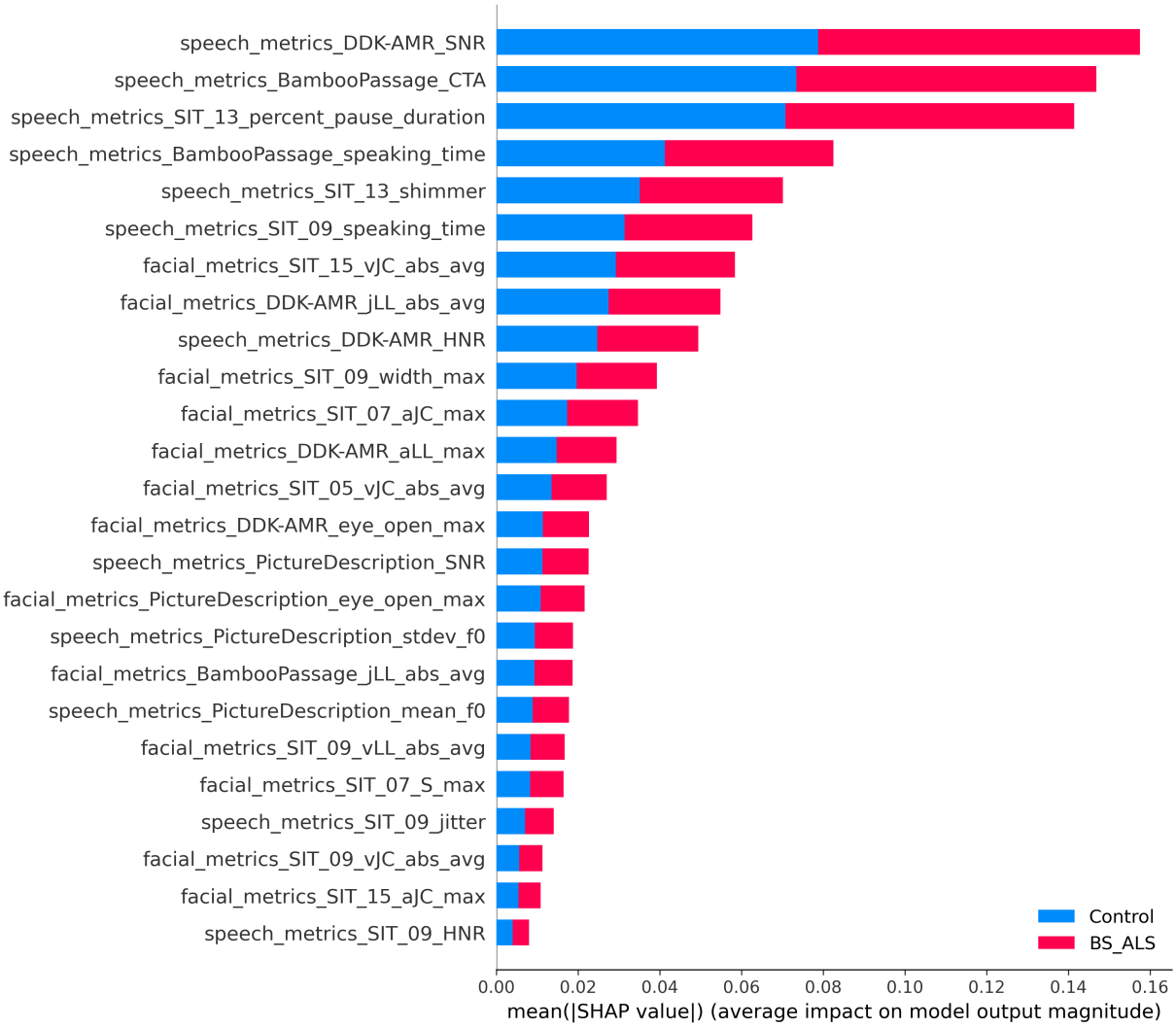



Figure 14: Shapley values: Controls versus BS ALS.

task) and speaking time (Bamboo, SIT 9) as well as the SNR in the DDK task. As can be seen in Figure 14, these features demonstrate a high contribution to the classifiers prediction. More specifically, SNR, CTA and PPT lead in feature contribution with a significant margin before other features, which is in line with their effect size magnitude compared to other features shown in Figure 9 in Section 6.3 and reflected a low rank sum shown in Section 6.4.1.

**Controls vs. BP ALS** None of the classification approaches was able to effectively learn to distinguish between BP ALS and controls. The performance of speech, facial and combined modalities including our feature selection pipeline is shown in Section 6.4.1.

| | Speech | Facial | Speech & Facial |
|---|---|---|---|
| Sensitivity | 0.53 | 0.51 | 0.52 |
| Specificity | 0.54 | 0.51 | 0.53 |
| AUC | 0.54 | 0.51 | 0.53 |
| F1 | 0.54 | 0.51 | 0.52 |

Table 18: Performance Metrics: BP ALS cases versus controls.

In BP ALS, there is only a single feature cluster that is selected in $>= 85\%$ of classification folds, which is the facial category *mouth symmetry*. The facial feature of average mouth symmetry ratio in the picture description task is also shown to be the most important feature for the best performing MLP classifier, which can be seen in Figure 15. In addition, this feature which shows the overall highest effect size between BP ALS cases and all controls (-0.41), indicating a lower symmetry ratio for BP ALS patients compared to controls. In addition, the Shapley analysis indicates a high mean average contribution of maximum eye opening (PicDesc), followed by the energy feature SNR in the SIT 9 task. However, while we also find small effect sizes for these features, our repeated classification experiment does not indicate consistency of selecting features from those categories.

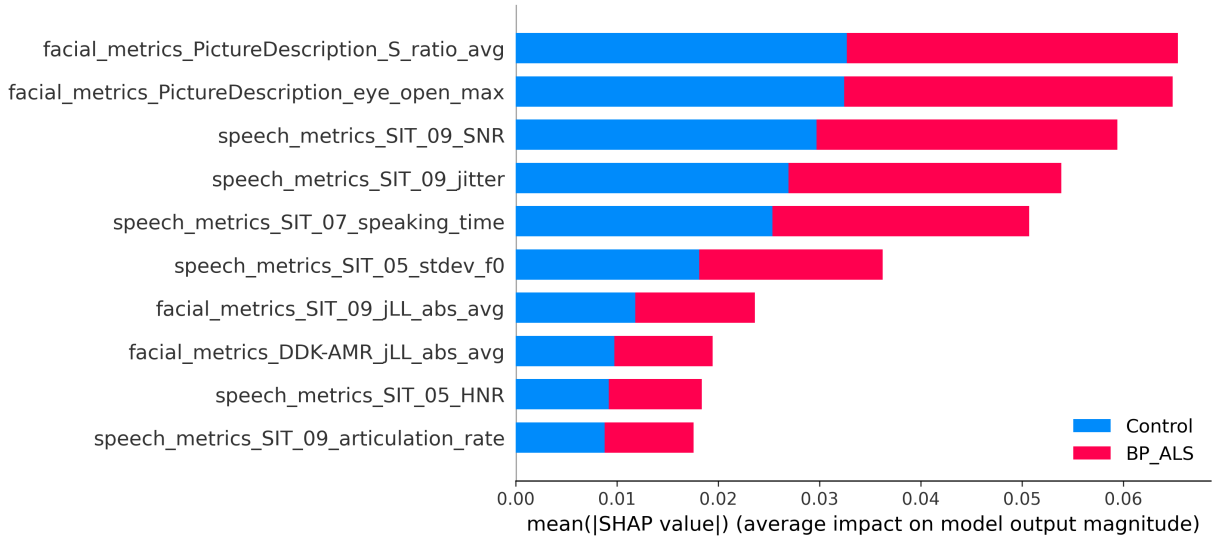



Figure 15: Shapley values: Controls versus BP ALS.

**Across disorders** To explore unique characteristics and potential overlaps among disorders within and across domains, we ran experiments within the mental and the neurological domain, and one classification across these domains. More specifically, we investigate depression vs.

schizophrenia, BP vs. BS ALS and schizophrenia vs. BS ALS. We selected the latter for the across-domain experiment because the binary case-control classification experiments demonstrate the highest discrimination power, while the feature importance analysis indicates various feature categories of similar importance for identifying schizophrenia and BS ALS.
As shown in Table 19, as in the control versus cases with a disorder experiments, the speech modality is more informative on its own than the facial modality, which is shown consistently across experiments. However, in both experiments involving schizophrenia, combining speech with facial information slightly increases the classification performance. Overall, the results indicate a lower discrimination power in these experiments compared to the best performances in distinguishing controls from cases with a disorder.

| Cohort | Speech | Facial | Speech & Facial |
|---|---|---|---|
| BS vs. BP ALS | **0.77** | 0.61 | **0.77** |
| Schizophrenia (SCHIZ) vs. Depression (DEP) | 0.77 | 0.65 | **0.78** |
| SCHIZ vs. BS ALS | 0.77 | 0.64 | **0.78** |

Table 19: F1-scores for across-disorder binary classification experiments.

The feature contributions shown in Figure 16 indicate a high importance of timing, voice quality and DDK-specific categories. Notably, two DDK features stand out in these analyses as they are both ranked among the top 3 features in both experiments, which is the facial feature *average lower lip jerk* and the speech feature *syllable rate*.
Regarding the feature contribution for distinguishing between depression and schizophrenia, we primarily find energy, timing and voice quality related features.
While features related to energy and timing categories are important in distinguishing both schizophrenia and BS ALS from controls, only one feature (PPT in the Bamboo reading task) ranks relatively high in importance for classifying between schizophrenia and BS ALS. However, according to the Shapley analysis, the classifier primarily focuses on voice quality, frequency features, and the facial feature *eye opening*. In particular, the voice quality feature *shimmer* ranks among the top three features twice (for SIT 7 & DDK task). Notably, the *shimmer* feature in the SIT 7 task holds the highest contribution margin among the top three features, emphasizing its substantial significance.

### 6.4.2 Multiclass Classification

**Performance** For the most complex task, discriminating controls and any individual disorder (5-class classification), using both speech and facial features, we obtain the best overall performance

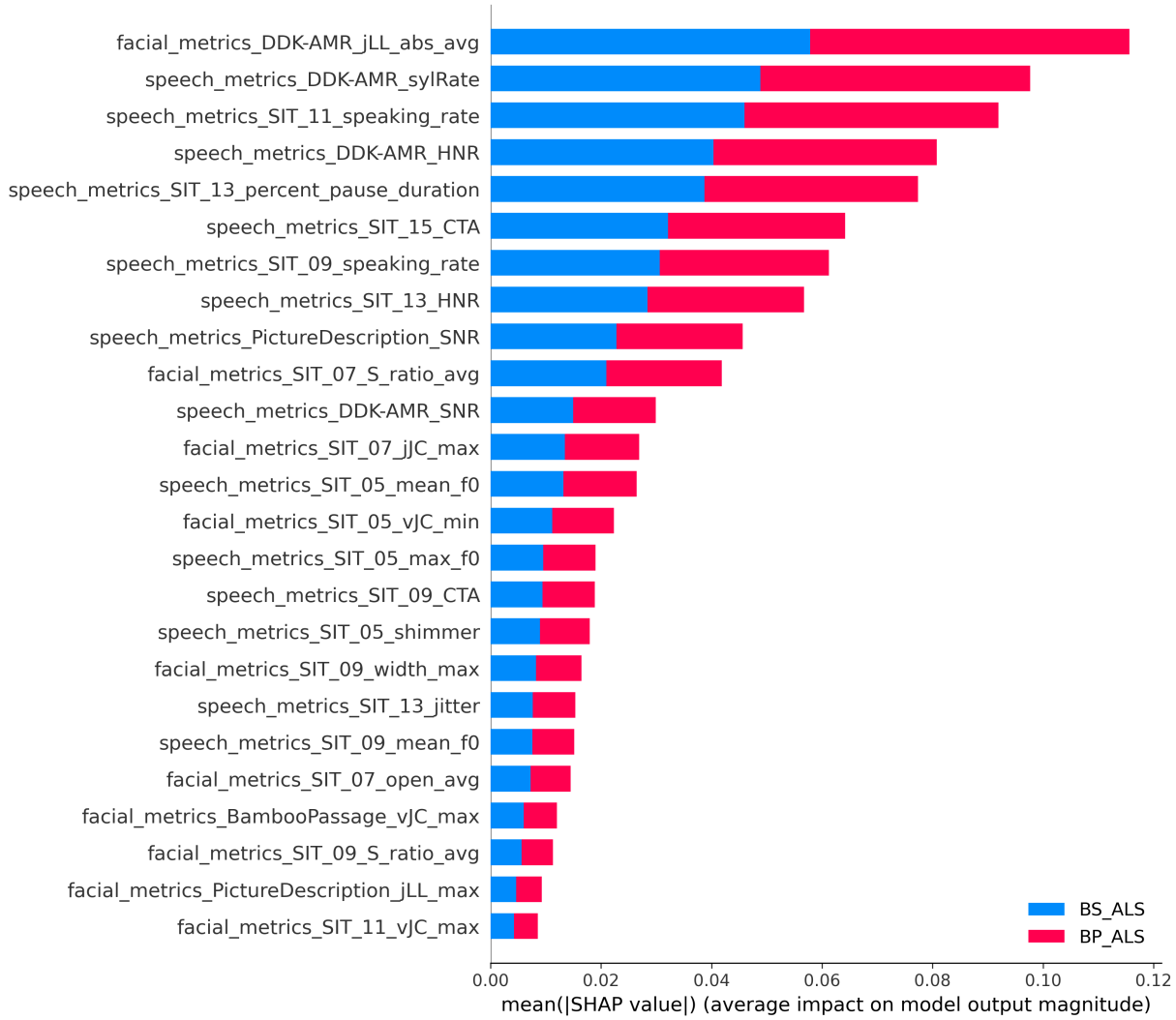



Figure 16: Shapley values: BP versus BS ALS.

with the MLP (F1-score: 0.53), as shown in Table 21. The MLP that implemented our proposed feature selection pipeline also performed better than all baseline models. Among baseline models, the embedded feature selection baseline achieved the highest performance (F1-score: 0.52).
As shown in Section 6.4.2, for detecting most cohorts correctly, namely controls, schizophrenia, and BS ALS, the per class F1-score is highest when combining speech and facial features. For detecting BP ALS, there is no performance difference between using only speech or speech and facial features. For depression, we achieve the best performance with speech features only.
Section 6.4.2 shows a confusion matrix that indicates the percentage of correct class predictions and with which they were confused. The model was most confident in detecting schizophrenia (75.42%), followed by BS ALS (64.17%).

| Cohort | Speech | Facial | Speech + Facial |
|---|---|---|---|
| Control | 0.33 | 0.28 | **0.34** |
| SCHIZ | 0.71 | 0.52 | **0.75** |
| BP ALS | **0.38** | 0.31 | **0.38** |
| BS ALS | 0.62 | 0.43 | **0.64** |
| DEP | **0.55** | 0.36 | 0.53 |
| Average | 0.52 | 0.38 | **0.53** |

Table 20: F1-scores per cohort for multi-class classification between all investigated disorders.

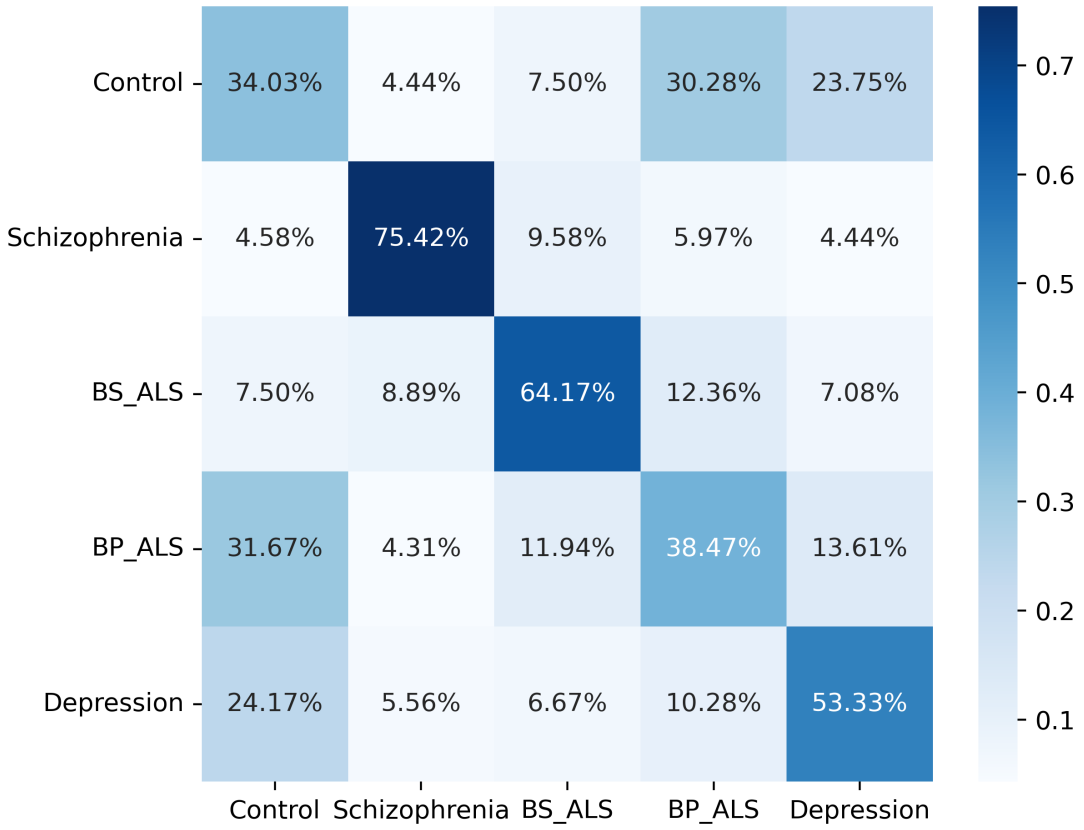



Figure 17: Normalized confusion matrix for 5-class classification. The x-axis shows the true labels, the y-axis the predicted ones.

Depression was correctly classified in 53.33%, which is well above chance in a 5-class classification. The model struggled with accurately predicting BP ALS (38.47%) and controls (34.03%).

Control subjects were most frequently mistaken for BP ALS (30.28%) and depression (23.75%) cases, and vice versa. Schizophrenic patients were least often confused with other cohorts. Among the cases of BS ALS, the most common, though rather infrequent, confusion occurred with BP

patients (12.36%). Although the model had difficulty effectively distinguishing between BP ALS and controls, the fact that BS ALS cases were most often misclassified as BP ALS cases may suggest that the model was able to capture similarities between BP ALS and BS ALS to some degree.

In addition, we conducted domain specific multiclass experiments, i.e. 3-class classification including controls and, respectively, mental or neurological disorders. We found trends and performance magnitudes to be in agreement with the ones in the five-class classification experiments, as shown in Table 21. Hence, for clarity, we will focus on the 5-class classification experiments in this section.

| Cohorts | Classes | FS + LR | FS + RF | FS + MLP | Filter BL | Wrapper BL | EMB BL |
|---|---|---|---|---|---|---|---|
| All | 5 | 0.51 | 0.52 | **0.53** | 0.51 | 0.48 | 0.52 |
| CONT vs. NEURO | 3 | **0.58** | 0.57 | **0.58** | 0.56 | **0.58** | 0.57 |
| CONT vs. MENTAL | 3 | 0.6 | **0.63** | 0.62 | 0.62 | 0.59 | 0.6 |

Table 21: Multi-class classification results for baselines and the feature selection selection pipeline assessed as F1 scores.

**Feature selection & importance** For the sake of consistency and due to overall good performance, we evaluated feature selection based on the best MLP model. As shown in Figure 18, in the multiclass classification task that included all cohorts, the Shapley analysis identified DDK-specific energy and articulation skills, such as cTV, SNR and syllable rate, as well as the voice quality feature shimmer, and the timing feature PPT among the top five most important features. More specifically, cTV (DDK task), stood out as the most important feature by demonstrating a relatively large margin in terms of its contribution to the classification outcome. Furthermore, the class-specific Shapley values indicate that this feature played a critical role in differentiating BS ALS cases from other cohorts.

In classifying BP ALS, average jerk of the lower lip in the DDK task showed a large contribution compared to other features.

For schizophrenia, speaking rate in the Bamboo reading task stood out as a feature that contributed relatively high to the prediction of schizophrenia compared with other cohorts. However, features associated with schizophrenia received relatively low Shapley values compared with other disorders. Note that this does not necessarily imply a lack of importance or relevance. Instead, it suggests that schizophrenia may have a more complex and multifaceted presentation, with multiple features collectively contributing to its identification. In light of the good classification performance in schizophrenia, the findings suggest that numerous features provide equally valu-

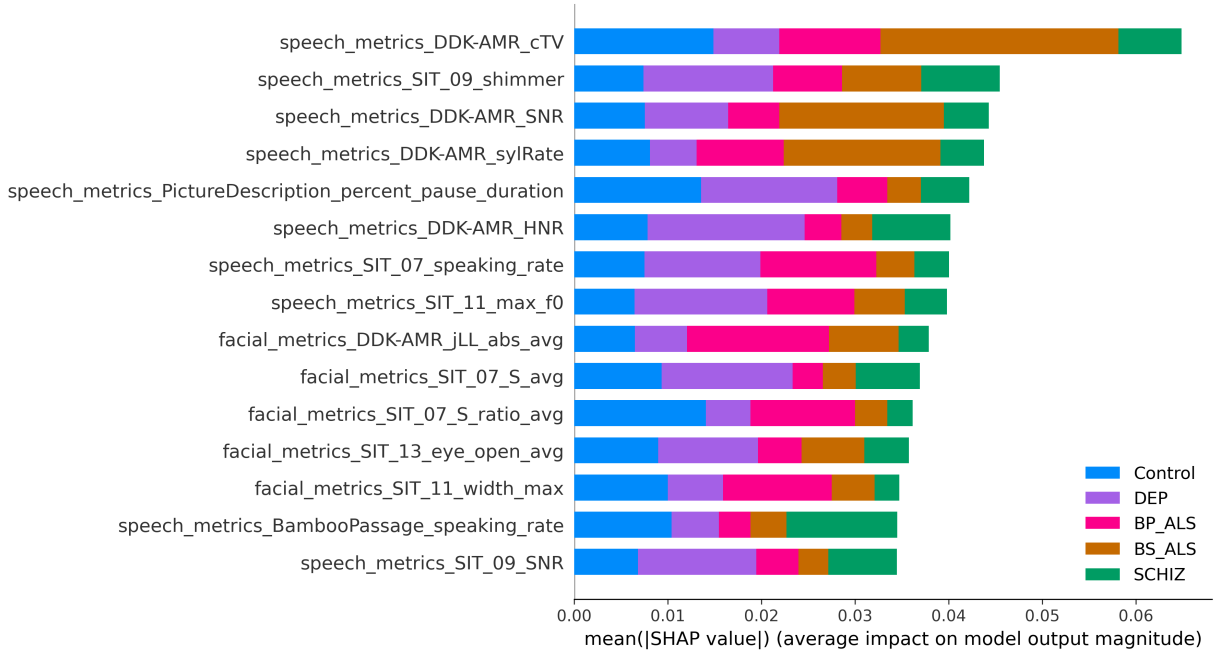



Figure 18: Shapley values for best performing multi-class model including all cohorts. F1-score: 0.56

able and complementary information for accurately classifying schizophrenia.

In the case of depression, voice quality, timing and energy-related features such as shimmer, pauses, and SNR were found to contribute significantly to the prediction of depression.

For predicting controls, the Shapley analysis did not suggest any specific features that were clearly contributing most to the prediction of this class compared to other cohorts.

# 7 Discussion

**Age correction and sex normalization** Age detrending and sex normalization are crucial steps in feature analysis to account for the distinct age-related trends observed in males and females, as well as differences between the sexes. Failing to consider these variables adequately may introduce bias and artificially induce differences between groups that are influenced more by demographic factors than the specific disorder under investigation.

Our analysis has shown that numerous features change with age. Age trends of medium effect size magnitudes are both shown for speech and facial features. Among speech features, timing metrics showed the most pronounced effects in males and a similar magnitude of change in females. However, it is critical to note that males and females show partially different age-related trends, which is shown in terms of the direction and magnitude of effect size in certain metrics. For example, opposite effects are observed for fundamental frequency (F0), with women having lower pitched voices with increasing age, whereas men tend to have slightly higher pitched voices. To account for these different trends, it is essential to perform age detrending separately for males and females.

Linear age correction may not be the optimal modelling approach, as age-related changes in speech and facial features may not always follow a strictly linear trajectory. In addition, physiological age does not necessarily match chronological age, as humans age at different rates due to numerous factors. However, because of its simplicity, age correction based on linear regression coefficients offers a straightforward approach that is less prone to overfitting compared to more complex, non-linear approaches. It captures the overall trend of age-related changes without being overly adjusted for the idiosyncrasies and noise present in the data.

Sex differences manifest more prominently in speech features than in facial features. Among speech features, pitch features exhibit high effect sizes, followed by voice quality and timing features.

The study design required sex to be reported as either male or female. We acknowledge that this may have unintentionally excluded individuals that were not comfortable sharing their sex at birth. Challenges associated with inclusion in terms of sexual identity need to be approached in research studies that require information about participants' sex or gender. Hence, future studies should target to include a broader range of populations, consider different sexual identities, and explore the complex interplay of age, sex, *and* sexual identity for feature analysis. In addition, future studies should consider further variables such as education, socioeconomic status and ethnicity, which was out of scope in this thesis.

**Redundancy analysis** Hierarchical clustering, as deployed in our analysis, is a sensible approach as it allows insight into the complex relationships between variables and clusters. The redundancy analysis identified reasonable clusters in speech and facial modalities. Metrics of the same or related domains grouped together, such as shimmer, jitter and HNR, all of which represent aspects of voice quality. Moreover, clusters were mostly coherent on the task-level, i.e. related tasks that may capture similar abilities in a certain metric, such as SIT and Bamboo reading tasks are most frequently clustered together, while metrics in the DDK task are frequently forming their own cluster as they capture very specific abilities.

Although most of the identified clusters represent reasonable groups, there are a few results, that require further investigation specifically clusters 6 and 7. These clusters, which include timing measures for reading tasks, show that SIT 5 and 9 are grouped separately from all other reading passages. This finding is inconclusive because SIT 7, which has a sentence length between 5 and 9, is clustered with longer reading passages than SIT 9. There could be several possible explanations for this, such as sentences in SIT 5 and 9 having more similarities in terms of the number of syllables, since sentence length in this task is defined based on the number of words. Alternatively, SIT 5 and 9 might share different levels of complexity, such as the rarity of words or difficulty of pronunciation. These factors should be further investigated.

Determining the optimal number of clusters per domain in redundancy analysis comes with the complex task of finding a balance between minimizing redundancy and capturing potentially subtle, yet crucial information. Future research should prioritize determining this optimal trade-off. Related to that, while we determined the distance threshold in a knowledge-driven manner, this can also be done in a data-driven way (Ienco and Meo, 2008), which would fully automate the employed clustering algorithm and avoid human introduced bias. Additionally, comparing the results with a completely knowledge-driven approach for grouping features would be an interesting direction to pursue.

**Classification experiments** Our analysis demonstrates that the approach that does not handle redundancy at all, the filter feature selection baseline, is not able to surpass other models in any of the evaluation metrics in any disorder. This indicates that considering feature redundancy plays a crucial role in improving performance. Notably, our feature selection approach, despite not focusing on maximizing classification performance, demonstrates similar performance to the wrapper method, where features are selected solely based on performance. However, we found overall only small differences among all the investigated combinations of feature selection methods and ML models, which were very specific to certain evaluation metrics and disorders. This suggests that the performance may be most constrained by the informativeness of the features

themselves.

In both BS ALS and schizophrenia, our analysis shows that these cohorts can be well distinguished with high sensitivity and specificity from healthy controls using speech features alone. In depression, a more comprehensive assessment including the evaluation of facial behavior is important, which is indicated by a higher classification performance for combined speech and facial modalities and high ranking facial features in feature importance. Still, the performance is lower than for schizophrenia or BS ALS. All classifiers, independent of the feature selection method, struggled with differentiating BP ALS cases from controls.

Like in our binary classification approach, we achieved good performance in detecting cases of schizophrenia and BS ALS across all multiclass classification experiments. Similarly, the classification of depression surpasses chance levels although showing overall lower performance compared to detecting BS ALS or schizophrenia. For BP ALS, as before, we consistently found random chance results.

The model most commonly confused BP ALS and depression with healthy controls. This is in line with our effect size analysis, where we found the fewest and smallest effects between controls and BP ALS, followed by depression. These observations suggest that speech and facial behavior of individuals with depressive symptoms and BP ALS, as captured by the features used in our analysis, may closely resemble that of healthy control subjects.

Notably, in case of misidentifying BS ALS cases, the classifier most frequently categorized them as BP ALS. Although distinguishing BP ALS cases from controls is challenging, this outcome indicates that the classifier may be able to capture condition-specific information from features that are shared across different stages of ALS, which may have led to this confusion.

While the investigated features are easily explainable to non-experts, less interpretable ones, such as log mel spectrograms or Mel Frequency Cepstral Coefficients (MFCCs) in the speech modality may be able to capture more nuanced and complex patterns in the data. Or, deep learning approaches for representation learning could be applied, such as Res-Net 50 (Li and Lima, 2021) in the facial modality. While such features can be powerful in capturing subtle details and nuances of audiovisual behavior, the inner workings of the deep learning model are not easily explainable or interpretable by non-experts. Moreover, deep learning models typically require a large amount of data for training in order to generalize well, which the datasets under investigation currently do not provide.

In sum, more sophisticated features may offer an advantage in assessing the fine differences in less salient speech and facial behavior in some diseases or early stages of a disease, but come with the mentioned limitations.

**Feature analysis** Our study demonstrates both speech and facial features to be useful for distinguishing between the cohorts under investigation. While speech features alone typically outperformed facial features alone, combined information from both modalities commonly enhanced performance. However, in some cases speech features alone performed superior to combined modalities, which indicates that facial features added complexity to the task while not providing a useful information gain. This was highly dependent on the specific task, in particular with respect to cohort combinations and classification set ups (binary vs. multiclass). For example, in distinguishing depression cases from controls, combined modalities lead to a better performance than single modality models. However, in the multiclass experiment, the accuracy of detecting depression decreased with adding facial information compared to speech features alone. Notably, all other cohorts that, in contrast to depression, did not demonstrate this in binary classification tasks, benefited from using both modalities. We attribute this to the fact that the interplay between speech and facial modalities and shared characteristics of disorders are intertwined, and more research is needed to better understand such patterns.

However, even when facial features do not seem to provide useful information, our feature analysis suggests facial information may still be valuable in such cases. Regarding the clinical usefulness of features, it is crucial to identify true disease markers rather than idiosyncratic characteristics. When there are only a few useful markers, the classification performance may be misleading, as it might result in the assumption that none of the features are valuable if the performance is poor. However, it is important to consider that some features could still be highly relevant but might require additional information to be effective in a classification experiment. In BP ALS, for example, it may be hard to find disorder-relevant information as patients are still in an early stage of the disorder, where speech impairment and other disabilities have not developed yet. However, we demonstrated the importance for a single facial feature concerning mouth symmetry. Since differences between the healthy population may be subtle, this one feature is not sufficient to diagnose ALS. Still, it is an important starting point and may indicate differences in face characteristics that capture symmetric behavior in some way. In agreement with our findings, Guarin et al. (2022) found an association between symmetry features (lip and mouth) and ALS severity assessed by the ALSFRS-R bulbar subscore while also reporting negative classification results for these features.

In BS ALS, we identified most salient feature contributions for various aspects of the timing domain, demonstrated in a higher PPT and speaking time as well as lower CTA. This is expected as BS ALS patients experience speech impairment due to the associated deterioration of bulbar motor function. Or findings are in line with previous research (Green et al., 2004; 2013).

In depression, studies have consistently shown that individuals with depression make longer pauses compared to healthy controls (Åsa Nilsonne, 1987; Cannizzaro et al., 2004; Mundt et al.,

2012) which is in line with our findings. This has been linked to psychomotor retardation (Hoffmann et al., 1985; Bennabi et al., 2013). Numerous studies have reported evidence for a slower speaking rate in depression (Darby and Hollien, 1977; Godfrey and Knight, 1984; Hardy et al., 1984). However, our study presents contrasting findings, suggesting an increased speaking rate (SIT 7) in read speech as one of the most important features. This discrepancy could be attributed to the heterogeneous nature of depression and differences in treatment of the disorder, in particular medication.

The differences in findings may also be related to the nature of the tasks employed. In our study, we focused on features extracted from non-spontaneous speech. However, research by Alghowinem et al. (2013) indicated that spontaneous speech contains more relevant information about depressive characteristics compared to read speech. Furthermore, they found that features extracted from the beginning of each sentence in the reading task yielded better results than using full reading passages, suggesting that diagnosing depression may be more effective before subjects fully engage in the task. Investigating this further would be valuable in future research.

We also observed a smaller average lip width as an important feature, which may be associated with decreased emotional expressivity, as indicated by reduced smiling and increased frowning. These findings align with previous studies highlighting similar patterns of emotional expression in depression (Scherer et al., 2014). A diminished emotional expressiveness or a general lack of positive affect are commonly found in depression (Sorg et al., 2012). More generally speaking, changes in lip width, as demonstrated in our study, including a reduction in the width of mouth movements during speech or a less expressive mouth posture, have been linked to depressive symptoms (Cohn et al., 2009).

In schizophrenia, we identified timing and articulation, voice quality and energy features to be most important when compared to healthy controls. Shimmer and HNR are indicative of different aspects of voice quality and commonly found across ML and clinical research (Zhao et al., 2022). In addition, shimmer has been demonstrated to have a stable negative correlation with negative symptoms present in schizophrenia such as blunted affect and alogia (Zhao et al., 2022).

We demonstrated that distinct mental and neurological disorders, depression, schizophrenia and two stages of disease progression in ALS, share characteristics, in particular, in speech production. These include timing-related features (such as speaking time and pauses) as well as energy related and articulatory features.

Despite the identified commonalities, features combine differently and hence, form different patterns. In schizophrenia, voice quality features such as shimmer and the HNR added crucial information for the classifier to distinguish schizophrenia cases from controls. Furthermore, the Shapley analysis of our multiclass experiments shows that the feature contribution patterns change when confronted with the more complex task of differential diagnosis. In binary classification

tasks, we found that the models typically focus on a small set of features for the prediction, as indicated by the Shapley curves. In contrast, in multi classification tasks, we observed a more balanced contribution among a larger set of features, indicating that the model needs to incorporate more complex information for the prediction.
Note that the feature analysis does not establish a straightforward causal relationship, but rather shows salient associations between specific features and cohorts. In addition, it is largely infeasible to obtain a complete picture of an individual's medical history and potentially co-occurring disorders that may affect an individual's health and thereby the assessments. Hence, the lack of comprehensive information induces uncertainty in the results. The use of large data sets in the future contributes to mitigating this limitation.

# 8 Summary

In this thesis, we presented a transparent feature selection pipeline that addresses demographic-specific biases, feature redundancy and provides interpretability of ML models.
Our demographics analysis demonstrated that accounting for age and sex is crucial for avoiding bias as we revealed age effects across various domains such as timing in speech features or mouth opening in facial features. Age-related changes differed for various metrics by sex. In addition, numerous features, in particular of the speech domain, are shown to differ between males and females. The redundancy analysis provided sensible clusters by grouping features of similar domains and tasks together. We showed that our proposed approach can be effectively applied in binary as well as more complex multiclass classification tasks. Our pipeline demonstrated high sensitivity and specificity in distinguishing controls from schizophrenia and BS ALS cases in particular. Our study revealed the following findings for specific diseases compared to controls:

- We find consistent evidence for voice quality, energy and timing features in schizophrenia.
- In BS ALS, we determine CTA, PPT and SNR among the most important features.
- In BP ALS, our analyses only reveal the facial feature of average mouth symmetry.
- In depression, we find strong feature contributions of eye opening, lip width, timing and DDK-related energy and articulation skills.

In general, our study suggests that both speech and facial features are useful for distinguishing between the cohorts under investigation. Our thorough feature investigation demonstrated that distinct mental and neurological disorders share characteristics, in particular, in speech production. These include predominantly timing-related features such as speaking time and pauses.
In this context, the across-disorder experiments have shown that the algorithm can discriminate disorders of the same domain with high accuracy despite common characteristics. In relation to this, we demonstrated the usefulness of our approach for the more complex task of differential diagnosis, as it worked well in multiclass classification, outperforming baseline approaches. Moreover, these analyses provided valuable insights into which features, relative to other disorders, are more disorder-specific, as opposed to features that are highly sensitive across multiple disorders. For schizophrenia, for example, we identify the voice quality feature of shimmer as important across experiments. In addition, we found that in the multiclass task, features contributed more uniformly compared to the binary classification experiments. This indicates that the model is able to detect unique patterns in this more complex task instead of focusing on a few very prominent features,

as shown in the binary experiments. To enhance understanding of feature contributions, future research should focus on examining the intertwined relationships between related disorders and controls and feature patterns.

# 9 References


Shumaila Aleem, Noor ul Huda, Rashid Amin, Samina Khalid, Sultan S. Alshamrani, and Abdullah Alshehri. Machine learning algorithms for depression: Diagnosis, insights, and research directions. *Electronics*, 11(7), 2022. ISSN 2079-9292. doi: 10.3390/electronics11071111. URL https://www.mdpi.com/2079-9292/11/7/1111.

Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Michael Breakspear, and Gordon Parker. Detecting depression: A comparison between spontaneous and read speech. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7547–7551, 2013. doi: 10.1109/ICASSP.2013.6639130.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Anton Batliner, Michael Neumann, Felix Burkhardt, Alice Baird, Sarina Meyer, Thang Vu, and Björn Schuller. Ethical awareness in paralinguistics: A taxonomy of applications. *International Journal of Human-Computer Interaction*, pages 1–18, 2022. doi: 10.1080/10447318.2022.2140385.

Roberto Battiti. Using mutual information for selecting features in supervised neural net learning. *Neural Networks, IEEE Transactions on*, 5:537 – 550, 1994. doi: 10.1109/72.298224.

V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann. Blazeface: Sub-millisecond neural face detection on mobile gpus. *CoRR*, abs/1907.05047, 2019. URL http://arxiv.org/abs/1907.05047.

Justin Bedo, Conrad Sanderson, and Adam Kowalczyk. An efficient alternative to svm based recursive feature elimination with applications in natural language processing and bioinformatics. In *Australian Conference on Artificial Intelligence*, volume 4304, pages 170–180. Springer, 2006.

Djamila Bennabi, Pierre Vandel, Charalambos Papaxanthis, Thierry Pozzo, and Emmanuel Haffen. Psychomotor retardation in depression: A systematic review of diagnostic, pathophysiologic, and therapeutic implications. *BioMed Research International*, 2013.

Visar Berisha, Krantsevich Chelsea, Hahn P Richard, Shira Hahn, Dasarathy Gautam, Turaga Pavan, and Julie Liss. Digital medicine and the curse of dimensionality. *NPJ Digital Medicine*, 4(1), 2021.

Visar Berisha, Chelsea Krantsevich, Gabriela Stegmann, Shira Hahn, and Julie Liss. Are reported accuracies in the clinical speech machine learning literature overoptimistic? In *Proc. Interspeech 2022*, pages 2453–2457, 2022. doi: 10.21437/Interspeech.2022-691.

Paul Boersma and Vincent Van Heuven. Speak and unspeak with praat. *Glot International*, 5 (9/10):341–347, 2001.

Andrea Bommert, Xudong Sun, Bernd Bischl, Jörg Rahnenführer, and Michel Lang. Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis*, 143:106839, 2020. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2019.106839. URL https://www.sciencedirect.com/science/article/pii/S016794731930194X.

Peter Buckley, Brian Miller, Douglas Lehrer, and David Castle. Psychiatric comorbidities and schizophrenia. *Schizophrenia bulletin*, 35:383–402, 12 2008. doi: 10.1093/schbul/sbn135.

Michael Cannizzaro, Brian Harel, Nicole Reilly, Phillip Chappell, and Peter Snyder. Voice acoustical measurement of the severity of major depression. *Brain and cognition*, 56:30–5, 2004. doi: 10.1016/j.bandc.2004.05.003.

Giovanni B Cassano, Stefano Pini, Marco Saettoni, Paola Rucci, and Liliana Dell'Osso. Occurrence and clinical correlates of psychiatric comorbidity in patients with psychotic disorders. *Journal of Clinical Psychiatry*, 59(2):60–68, 1998.

Jesse M. Cedarbaum, Nancy Stambler, Errol Malta, Cynthia Fuller, Dana Hilt, Barbara Thurmond, and Arline Nakanishi. The alsfrs-r: a revised als functional rating scale that incorporates assessments of respiratory function. *Journal of the Neurological Sciences*, 169(1):13–21, 1999. ISSN 0022-510X. doi: https://doi.org/10.1016/S0022-510X(99)00210-5. URL https://www.sciencedirect.com/science/article/pii/S0022510X99002105.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Publishers, Hillsdale, NJ, 2nd edition, 1988.

Jeffrey F. Cohn, T. S. Kruez, I. Matthews, Ying Yang, Minh Hoai Nguyen, Michael T. Padilla, Feng Zhou, and Fernando De la Torre. Detecting depression from facial actions and vocal

prosody. *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–7, 2009.

Nicholas Cummins, Jyoti Joshi, Abhinav Dhall, Vidhyasaharan Sethu, Roland Goecke, and Julien Epps. Diagnosis of depression by behavioural signals: A multimodal approach. In *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, AVEC '13, page 11–20, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450323956. doi: 10.1145/2512530.2512535. URL https://doi.org/10.1145/2512530.2512535.

John K. Darby and Harry Hollien. Vocal and speech patterns of depressive patients. *Folia phoniatrica*, 29(4):279–291, 1977. doi: 10.1159/000264098. URL https://doi.org/10.1159/000264098.

Kristin Diekmann, Magdalena Kuźma-Kozakiewicz, Maria Piotrkiewicz, Marta Gromicho, Julian Grosskreutz, Peter M. Andersen, Mamede de carvalho, Hilmi Uysal, Alma Osmanovic, Olivia Schreiber-Katz, Susanne Petri, and Sonja Körner. Impact of comorbidities and co-medication on disease onset and progression in a large german als patient group. *Journal of Neurology*, 267, 07 2020. doi: 10.1007/s00415-020-09799-z.

Ulf Dimberg and Lars-Olov Lundquist. Gender differences in facial reactions to facial expressions. *Biological Psychology*, 30(2):151–159, 1990. ISSN 0301-0511. doi: https://doi.org/10.1016/0301-0511(90)90024-Q. URL https://www.sciencedirect.com/science/article/pii/030105119090024Q.

Hanna Drimalla, Tobias Scheffer, Niels Landwehr, Irina Baskow, Stefan Roepke, Behnoush Behnia, and Isabel Dziobek. Towards the automatic detection of social biomarkers in autism spectrum disorder: introducing the simulated interaction task (sit). *npj Digital Medicine*, 3:25, 2020. doi: 10.1038/s41746-020-0227-5.

Farshad Falahati, Daniel Ferreira, J-Sebastian Muehlboeck, Hilkka Soininen, Patrizia Mecocci, Bruno Vellas, Magdalini Tsolaki, Iwona Kłoszewska, Christian Spenger, Simon Lovestone, Maria Eriksdotter, Lars-Olof Wahlund, Andy Simmons, and Eric Westman. The effect of age correction on multivariate classification in alzheimer's disease, with a focus on the characteristics of incorrectly and correctly classified subjects. *Brain Topography*, In-press, 03 2016. doi: 10.1007/s10548-015-0455-1.

Valery Feigin, Emma Nichols, Tahiya Alam, Marlena Bannick, Ettore Beghi, Natacha Blake, William Culpepper, E. Dorsey, Alexis Elbaz, Richard Ellenbogen, James Fisher, Christina

Fitzmaurice, Giorgia Giussani, Linda Glennie, Spencer James, Catherine Johnson, Nicholas Kassebaum, Giancarlo Logroscino, Benoît Marin, and Theo Vos. Global, regional, and national burden of neurological disorders, 1990-2016: a systematic analysis for the global burden of disease study 2016. *The Lancet Neurology*, 18:459–480, 05 2019. doi: 10.1016/S1474-4422(18) 30499-X.

Wolfgang Gaebel and Wolfgang Woelwer. Facial expression in the course of schizophrenia and depression. *European archives of psychiatry and clinical neuroscience*, 254:335–42, 2004. doi: 10.1007/s00406-004-0510-5.

Manosij Ghosh, Ritam Guha, Ram Sarkar, and Ajith Abraham. A wrapper-filter feature selection technique based on ant colony optimization. *Neural Comput. Appl.*, 32(12): 7839–7857, 2020. ISSN 0941-0643. doi: 10.1007/s00521-019-04171-3. URL https://doi.org/10.1007/s00521-019-04171-3.

Hamish P. D. Godfrey and Robert G. Knight. The validity of actometer and speech activity measures in the assessment of depressed patients. *The British Journal of Psychiatry*, 145(2): 159–163, 1984. doi: 10.1192/bjp.145.2.159.

Pablo Granitto, Cesare Furlanello, Franco Biasioli, and Flavia Gasperi. Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems - CHEMOMETR INTELL LAB SYST*, 83, 2006. doi: 10.1016/j.chemolab.2006.01.007.

Alan I Green, Carla M Canuso, Mark J Brenner, and Joanne D Wojcik. Detection and management of comorbidity in patients with schizophrenia. *Psychiatric Clinics*, 26(1):115–139, 2003.

Jordan Green, Yana Yunusova, Mili Kuruvilla, Jun Wang, Gary Pattee, Lori Synhorst, Lorne Zinman, and James Berry. Bulbar and speech motor assessment in als: Challenges and future directions. *Amyotrophic lateral sclerosis frontotemporal degeneration*, 07 2013. doi: 10.3109/ 21678421.2013.817585.

Jordan R. Green, David R. Beukelman, and Lauren J. Ball. Algorithmic estimation of pauses in extended speech samples of dysarthric and typical speech. *Journal of Medical Speech-Language Pathology*, 12:149–154, 2004.

Diego Guarin, Babak Taati, Agessandro Abrahão, Lorne Zinman, and Yana Yunusova. Video-based facial movement analysis in the assessment of bulbar amyotrophic lateral sclerosis: Clinical validation. *Journal of Speech, Language, and Hearing Research*, 65:1–12, 11 2022. doi: 10.1044/2022_JSLHR-22-00072.

Jean-Philippe Guilloux, Marianne Seney, Nicole Edgar, and Etienne Sibille. Integrated behavioral z-scoring increases the sensitivity and reliability of behavioral phenotyping in mice: Relevance to emotionality and sex. *Journal of neuroscience methods*, 197:21–31, 2011. doi: 10.1016/j.jneumeth.2011.01.019.

Blaise Hanczar, Jianping Hua, Chao Sima, John Weinstein, Michael Bittner, and Edward Dougherty. Small-sample precision of roc-related estimates. *Bioinformatics (Oxford, England)*, 26:822–30, 2010. doi: 10.1093/bioinformatics/btq037.

Patrick Hardy, Roland Jouvent, and Daniel Widlöcher. Speech pause time and the retardation rating scale for depression (erd): Towards a reciprocal validation. *Journal of Affective Disorders*, 6(1):123–127, 1984. ISSN 0165-0327. doi: https://doi.org/10.1016/0165-0327(84)90014-4. URL https://www.sciencedirect.com/science/article/pii/0165032784900144.

Mohammad Ehsan Heidari, Jamshid Nadali, Amir Parouhan, Mohammad Azarafraz, Seyed Mahmoud Tabatabai, Seyed Saeed Naghibi Irvani, Fatemeh Eskandari, and Alireza Gharebaghi. Prevalence of depression among amyotrophic lateral sclerosis (als) patients: A systematic review and meta-analysis. *Journal of affective disorders*, 287:182–190, 2021. doi: 10.1016/j.jad.2021.03.015.

Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1, 1995. doi: 10.1109/ICDAR.1995.598994.

Guy M Hoffmann, Jean Claude Gonze, and Julien Mendlewicz. Speech pause time as a method for the evaluation of psychomotor retardation in depressive illness. *The British journal of psychiatry : the journal of mental science*, 146:535–538, 1985. doi: 10.1192/bjp.146.5.535. URL https://doi.org/10.1192/bjp.146.5.535.

Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, and Wojciech Samek. *Explainable AI Methods - A Brief Overview*, pages 13–38. Springer International Publishing, Cham, 2022. ISBN 978-3-031-04083-2. doi: 10.1007/978-3-031-04083-2_2. URL https://doi.org/10.1007/978-3-031-04083-2_2.

K.D. Hopkins and G.V. Glass. *Basic Statistics for the Behavioral Sciences*. Prentice-Hall, Englewood Cliffs, N.J., 1978.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by

masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291.

Eric Hunter, Mara Kapsner-Smith, Patrick Pead, Megan Engar, and Wesley Brown. Age and speech production: A 50-year longitudinal study. *Journal of the American Geriatrics Society*, 60:1175–7, 2012. doi: 10.1111/j.1532-5415.2012.03983.x.

Dino Ienco and Rosa Meo. Exploration and reduction of the feature space by hierarchical clustering. volume 2, pages 577–587, 04 2008. doi: 10.1137/1.9781611972788.53.

Haihua Jiang, Bin Hu, Zhenyu Liu, Lihua Yan, Tianyang Wang, Fei Liu, Huanyu Kang, and Xiaoyu Li. Investigation of different speech types and emotions for detecting depression using different classifiers. *Speech Communication*, 90:39–46, 2017. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2017.04.001. URL https://www.sciencedirect.com/science/article/pii/S0167639316303053.

Dr. Kala, Suganya Balamurugan, and Geetha Subbiah. Multi filtration feature selection (mffs) to improve the discriminatory ability in clinical data set. *Applied Computing and Informatics*, 31, 2014. doi: 10.1016/j.aci.2014.03.002.

Amandeep Kaur, Kalpna Guleria, and Naresh Kumar Trivedi. Feature selection in machine learning: Methods and comparison. In *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 789–795, 2021. doi: 10.1109/ICACITE51222.2021.9404623.

Hardik Kothare, Michael Neumann, Jackson Liscombe, Oliver Roesler, William Burke, Andrew Exner, Sandy Snyder, Andrew Cornish, Doug Habberstad, David Pautler, David Suendermann-Oeft, Jessica Huber, and Vikram Ramanarayanan. Statistical and clinical utility of multimodal dialogue-based speech and facial metrics for parkinson's disease assessment. pages 3658–3662, 2022. doi: 10.21437/Interspeech.2022-11048.

K. Kroenke, R.L. Spitzer, and Janet Williams. The phq-9 validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16, 09 2001. doi: 10.1046/j1525-1497200101600960 6x.

Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B.W. Williams, Joyce T. Berry, and Ali H. Mokdad. The phq-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, 114(1):163–173, 2009. ISSN 0165-0327. doi: https://doi.org/10.1016/j.jad.2008.06.026. URL https://www.sciencedirect.com/science/article/pii/S0165032708002826.

Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. SpringerLink : Bücher. Springer New York, 2013. ISBN 9781461468493. URL https://books.google.com/books?id=xYRDAAAAQBAJ.

S Körner, Katja Kollewe, J Ilsemann, Annika Karch, R Dengler, K Krampfl, and Susanne Petri. Prevalence and prognostic impact of comorbidities in amyotrophic lateral sclerosis. *European journal of neurology : the official journal of the European Federation of Neurological Societies*, 20, 10 2012. doi: 10.1111/ene.12015.

Vinh La, Sungyoung Lee, Young Tack Park, and Brian d'Auriol. A novel feature selection method based on normalized mutual information. *Applied Intelligence - APIN*, 37, 2012. doi: 10.1007/s10489-011-0315-y.

Adam Lammert, James Williamson, Austin Hess, Tejash Patel, Thomas Quatieri, Hui Liao, Alexander Lin, and Kristin Heaton. Noninvasive estimation of cognitive status in mild traumatic brain injury using speech production and facial expression. pages 105–110, 2017. doi: 10.1109/ACII.2017.8273586.

Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu, and Haifeng Wang. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. pages 1299–1308, 2017. doi: 10.18653/v1/D17-1134.

Bin Li and Dimas Lima. Facial expression recognition via resnet-50. *International Journal of Cognitive Computing in Engineering*, 2, 2021. doi: 10.1016/j.ijcce.2021.02.002.

Yuanchao Li, Tianyu Zhao, and Tatsuya Kawahara. Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning. pages 2803–2807, 2019. doi: 10.21437/Interspeech.2019-2594.

Xiaohui Lin, Fufang Yang, Lina Zhou, Peiyuan Yin, Hongwei Kong, Wenbin Xing, Xin Lu, Lewen Jia, Quancai Wang, and Guowang Xu. A support vector machine-recursive feature elimination feature selection method based on artificial contrast variables and mutual information. *Journal of chromatography B*, 910:149–155, 2012.

Dichao Liu, Yu Wang, Kenji Mase, and Jien Kato. Attention-based multi-task learning for fine-grained image classification. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1499–1503, 2021. doi: 10.1109/ICIP42928.2021.9506745.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Carol Zander Malatesta, C. E. Izard, Clayton Culver, and Mark J. Nicolich. Emotion communication skills in young, middle-aged, and older women. *Psychology and aging*, 2 2:193–203, 1987.

Scott Menard. Standardized regression coefficients. In Michael S. Lewis-Beck, Alan Bryman, and Tim Futing Liao, editors, *The Sage Encyclopedia of Social Science Research Methods*, pages 1069–1070. Sage Publications, Thousand Oaks, CA, USA, 2004. ISBN 9780761923633. doi: 10.4135/9781412950589.n959.

Junaid Muhammad, Sajid Ali, Fatma Eid, Shaker El-Sappagh, and Tamer Abuhmed. Explainable machine learning models based on multimodal time-series data for the early detection of parkinson's disease. *Computer Methods and Programs in Biomedicine*, 234, 2023. doi: 10.1016/j.cmpb.2023.107495.

James C Mundt, Adam P Vogel, Douglas E Feltner, and William R Lenderking. Vocal acoustic biomarkers of depression severity and treatment response. *Biological psychiatry*, 72(7):580–587, 2012.

Michael Neumann, Oliver Roesler, Jackson Liscombe, Hardik Kothare, David Suendermann-Oeft, J. D. Berry, E. Fraenkel, R. Norel, Aria Anvar, I. Navar, A. V. Sherman, Jordan R. Green, and Vikram Ramanarayanan. Multimodal dialog based speech and facial biomarkers capture differential disease progression rates for als remote patient monitoring. In *Proceedings of the 32nd International Symposium on Amyotrophic Lateral Sclerosis and Motor Neuron Disease*, Virtual, 2021.

Michal Novotny, Jan Melechovsky, Krystof Rozenstoks, Tereza Tykalova, Petr Kryze, Marek Kanok, Jiri Klempir, and Jan Rusz. Comparison of automated acoustic methods for oral diadochokinesis assessment in amyotrophic lateral sclerosis. *Journal of speech, language, and hearing research : JSLHR*, 63(10):3453–3460, 2020. doi: 10.1044/2020_JSLHR-20-00109.

Hanchuan Peng, Fuhui Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005. doi: 10.1109/TPAMI.2005.159.

Nicholas Pudjihartono, Tayaza Fadason, Andreas Kempa-Liehr, and Justin O'Sullivan. A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in Bioinformatics*, 2:927312, 2022. doi: 10.3389/fbinf.2022.927312.

Joerg Pueschel, Hans Stassen, G. Bomben, Christian Scharfetter, and Daniel Hell. Speaking behavior and speech sound characteristics in acute schizophrenia. *Journal of psychiatric research*, 32:89–97, 1998. doi: 10.1016/S0022-3956(98)00046-6.

Michael Raymer, Travis Doom, Leslie Kuhn, and William Punch. Knowledge discovery in medical and biological datasets using a hybrid bayes classifier/evolutionary algorithm. *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, 33:802–13, 2003. doi: 10.1109/TSMCB.2003.816922.

Vanessa Richter, Michael Neumann, Hardik Kothare, Oliver Roesler, Jackson Liscombe, David Suendermann-Oeft, Sebastian Prokop, Anzalee Khan, Christian Yavorsky, Jean-Pierre Lindenmayer, and Vikram Ramanarayanan. Towards multimodal dialog-based speech & facial biomarkers of schizophrenia. In *Companion Publication of the 2022 International Conference on Multimodal Interaction*, ICMI '22 Companion, page 171–176, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393898. doi: 10.1145/3536220.3558075. URL https://doi.org/10.1145/3536220.3558075.

Oliver Roesler, Hardik Kothare, William Burke, Michael Neumann, Jackson Liscombe, Andrew Cornish, Doug Habberstad, David Pautler, David Suendermann-Oeft, and Vikram Ramanarayanan. Exploring facial metric normalization for within- and between-subject comparisons in a multimodal health monitoring agent. In *Companion Publication of the 2022 International Conference on Multimodal Interaction*, ICMI '22 Companion, page 160–165, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393898. doi: 10.1145/3536220.3558071. URL https://doi.org/10.1145/3536220.3558071.

Stefan Scherer, Giota Stratou, Gale Lucas, Marwa Mahmoud, Jill Boberg, Jonathan Gratch, and Louis-Philippe Morency. Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image and Vision Computing*, 32(10):648–658, 2014.

Lloyd S. Shapley. *Notes on the N-Person Game – II: The Value of an N-Person Game*. RAND Corporation, Santa Monica, CA, 1951. doi: 10.7249/RM0670.

Andrew F. Siegel and Michael R. Wagner. Chapter 12 - multiple regression: Predicting one variable from several others. In Andrew F. Siegel and Michael R. Wagner, editors, *Practical Business Statistics (Eighth Edition)*, pages 371–431. Academic Press, eighth edition

edition, 2022. ISBN 978-0-12-820025-4. doi: https://doi.org/10.1016/B978-0-12-820025-4.00012-9. URL https://www.sciencedirect.com/science/article/pii/B9780128200254000129.

Alan K Silbergleit, Arthur F Johnson, and Barbara H Jacobson. Acoustic analysis of voice in individuals with amyotrophic lateral sclerosis and perceptually normal vocal quality. *Journal of Voice*, 11(2):222–231, 1997.

Sonja Sorg, Claus Vögele, Nadine Furka, and Andrea Meyer. Perseverative thinking in depression and anxiety. *Frontiers in Psychology*, 3, 2012. ISSN 1664-1078. doi: 10.3389/fpsyg.2012.00020. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2012.00020.

David Suendermann-Oeft, Amanda Robinson, Andrew Cornish, Doug Habberstad, David Pautler, Dirk Schnelle-Walka, Franziska Haller, Jackson Liscombe, Michael Neumann, Mike Merrill, Oliver Roesler, and Renko Geffarth. Nemsi: A multimodal dialog system for screening of neurological or mental conditions. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, IVA '19, page 245–247, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366724. doi: 10.1145/3308532.3329415. URL https://doi.org/10.1145/3308532.3329415.

Bogdan Tomik and Roberto J Guiloff. Dysarthria in amyotrophic lateral sclerosis: A review. *Amyotrophic Lateral Sclerosis*, 11(1-2):4–15, 2010.

Basil Varkey. Principles of clinical ethics and their application to practice. *Medical Principles and Practice*, 30, 2020. doi: 10.1159/000509119.

Yap Bee Wah, Nuhu Ibrahim, H.A. Hamid, Shuzlina Rahman, and Simon Fong. Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy. *Pertanika Journal of Science and Technology*, 26:329–340, 2018.

Joe H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244, 1963.

James Williamson, Elizabeth Godoy, Miriam Cha, Adrianne Schwarzentruber, Pooya Khorrami, Youngjune Gwon, Hsiang-Tsung Kung, Charlie Dagli, and Thomas Quatieri. Detecting depression using vocal, facial and semantic communication cues. pages 11–18, 2016. doi: 10.1145/2988257.2988263.

Juanying Xie and Chunxia Wang. Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases. *Expert Systems with Applications*, 38:5809–5815, 2011. doi: 10.1016/j.eswa.2010.10.050.

Qian Zhao, Wei-Qiang Wang, Hai-Zhen Fan, Dan Li, Yu-Jing Li, Yan-Ling Zhao, Zhi-Xin Tian, Zhi-Ren Wang, Yan-Ling Tan, and Shao-Ping Tan. Vocal acoustic features may be objective biomarkers of negative symptoms in schizophrenia: A cross-sectional study. *Schizophrenia Research*, 250:180–185, 2022. doi: 10.1016/j.schres.2022.11.013. URL https://doi.org/10.1016/j.schres.2022.11.013.

Åsa Nilsonne. Acoustic analysis of speech variables during depression and after improvement. *Acta Psychiatrica Scandinavica*, 76, 1987.

# List of Figures



# List of Tables

# Abbreviations

**ALS** Amyotrophic Lateral Sclerosis

**BP** Bulbar pre-symptomatic

**BL** Baseline

**BS** Bulbar symptomatic

**DEP** Depression

**DDK** Diadochokinetic

**ES** Effect Size

**FS** Feature Selection Pipeline

**LR** Logistic Regression

**ML** Machine Learning

**MWU** Mann-Whitney U

**MLP** Multilayer perceptron

**PicDesc** Picture Description Task

**RF** Random Forest

**SCHIZ** Schizophrenia