TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Erik Illaste 221678IAPM

DETECTION OF COGNITIVE DECLINE FROM SPONTANEOUS SPEECH: COMPARING MODEL PERFORMANCE AND INTERPRETABILITY

Master's Thesis

Supervisor: Tanel Alumäe
PhD

Tallinn 2025

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Erik Illaste 221678IAPM

KOGNITIIVSE LANGUSE TUVASTAMINE SPONTAANSEST KÕNEST: MUDELITE JÕUDLUSE JA TÕLGENDATAVUSE VÕRDLUS

Magistritöö

Juhendaja: Tanel Alumäe
PhD

Tallinn 2025

Author's Declaration of Originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Erik Illaste

12.05.2025

Abstract

Detection of cognitive decline, including conditions such as mild cognitive impairment and Alzheimer's dementia, is crucial for enabling timely intervention and supporting long-term care planning. Spontaneous speech is a promising medium for scalable, non-invasive cognitive screening, as impairments in language often emerge in the early stages of cognitive deterioration.

This thesis investigates the use of machine learning to detect cognitive decline from spontaneous speech, with an emphasis on both predictive performance and model interpretability. Models are trained on acoustic, text-based, and demographic features, combining interpretable features with representations derived from deep learning models. External corpora are explored alongside the primary dataset, with their use limited to cases where alignment in distribution and task relevance could be established.

The results demonstrate that deep learning-based acoustic features show a stronger potential for diagnostic classification tasks, effectively distinguishing between healthy individuals, those with mild impairment, and those with dementia. In contrast, interpretable features—such as fluency metrics, pause patterns, and demographic indicators—proved more effective for predicting cognitive assessment scores. A novel Joint Interpretability Index, combined with Pareto front analysis, reveals a clear trade-off: for classification, the most accurate models tend to be less interpretable, whereas in regression tasks, the most interpretable model–feature set combinations also achieve the best performance.

The thesis is written in English and is 77 pages long, including 5 chapters, 41 figures, and 23 tables.

Annotatsioon

Kognitiivse languse tuvastamine spontaansest kõnest: mudelite jõudluse ja tõlgendatavuse võrdlus

Kognitiivse languse, sealhulgas kerge kognitiivse häire ja Alzheimeri dementsuse, tuvastamine on oluline, et võimaldada õigeaegset sekkumist ja toetada pikaajalise hoolduse planeerimist. Spontaanne kõne on paljulubav vahend skaleeritavaks ja mitteinvasiivseks kognitiivseks sõeltestimiseks, kuna keelelised häired ilmnevad sageli kognitiivse halvenemise varajases staadiumis.

Käesolev magistritöö uurib masinõppe kasutamist kognitiivse languse tuvastamiseks spontaanse kõne põhjal, pöörates rõhku nii ennustustäpsusele kui ka mudelite tõlgendatavusele. Mudelid treenitakse akustiliste, tekstipõhiste ja demograafiliste tunnuste alusel, ühendades tõlgendatavad tunnused süvaõppe mudelitest saadud representatsioonidega. Lisaks põhiandmestikule uuritakse ka väliseid andmekogumeid, mille kasutamine piirdub juhtudega, kus on võimalik kindlaks teha jaotuslik vastavus ja ülesande asjakohasus.

Tulemused näitavad, et süvaõppel põhinevatel akustilistel tunnustel on suurem potentsiaal diagnostilistes klassifitseerimisülesannetes, eristades tõhusalt terveid isikuid, kerge kognitiivse häirega isikuid ja dementsusega patsiente. Seevastu tõlgendatavad tunnused — nagu soravusmõõdikud, pausimustrid ja demograafilised näitajad — osutusid tõhusamaks kognitiivsete testide tulemuste ennustamisel. Uus liittõlgendatavuse indeks koos Pareto-optimaalsuse analüüsiga toob esile selge kompromissi: klassifitseerimise puhul on kõige täpsemad mudelid tavaliselt vähem tõlgendatavad, samas kui regressiooniülesannetes saavutavad parima tulemuse need mudeli ja tunnuste kombinatsioonid, mis on ka kõige paremini tõlgendatavad.

Magistritöö on kirjutatud inglise keeles ja sisaldab teksti 77 leheküljel, 5 peatükki, 41 joonist ja 23 tabelit.

List of Abbreviations and Terms

AD	Alzheimer's dementia
CoT	Chain-of-thought
CTD	Cookie Theft Picture Description Task
CV	Cross-validation
DT	Decision Tree
EDA	Exploratory Data Analysis
eGeMAPS	Extended Geneva Minimalistic Acoustic Parameter Set
GB	Gradient Boosting
HC	Healthy Control
JII	Joint Interpretability Index
KNN	K-Nearest Neighbors
LIR	Linear Regression
LLM	Large Language Model
LR	Logistic Regression
MCI	Mild Cognitive Impairment
MFCCs	Mel-Frequency Cepstral Coefficients
ML	Machine Learning
MLP	Multilayer Perceptron
MMSE	Mini-Mental State Examination
MoCa	Montreal Cognitive Assessment
NB	Naive Bayes
PFT	Phonemic Fluency Task
RF	Random Forest
RMSE	Root Mean Squared Error
RR	Ridge Regression
SFT	Semantic Fluency Task
SHAP	SHapley Additive exPlanation
SVM	Support Vector Machine
t-SNE	t-distributed Stochastic Neighbor Embedding
VC	Voting Classifier
XGB	Extreme Gradient Boosting

Table of Contents

List of Figures

List of Tables

1. Introduction

Detection of cognitive decline, including mild cognitive impairment (MCI) and Alzheimer's dementia (AD), is critical for timely interventions that may slow disease progression and enhance quality of life. Conventional diagnostic methods, such as the Montreal Cognitive Assessment (MoCA) [1] and Mini-Mental State Examination (MMSE) [2], though reliable, are resource-intensive, limiting their scalability for large-scale or continuous monitoring. This highlights the need for automated, non-invasive tools for cognitive assessment.

Spontaneous speech analysis offers a promising approach to cognitive screening, as language deficits are among the earliest signs of cognitive decline. Characteristics like reduced fluency, lexical retrieval difficulties, and disorganized speech patterns are frequently observed in individuals with cognitive impairments. Machine learning (ML) models, leveraging both interpretable linguistic and acoustic features as well as latent representations learned from raw speech, can help detect these subtle changes.

While deep learning methods have demonstrated promising results in detecting cognitive decline, their lack of interpretability poses challenges for clinical use, where transparency in decision-making is key. Clinicians require models that not only perform well but are also interpretable to ensure reliable diagnoses. This research focuses primarily on ML models trained on combinations of interpretable features and learned features from pretrained deep learning models, investigating which features are most indicative of cognitive decline and how their integration impacts the trade-off between accuracy and interpretability.

1.1 Motivation

Alzheimer's dementia, along with other forms of dementia, is among the leading causes of death worldwide, accounting for approximately two million deaths annually (Figure 1). As the global population ages [3], there is a pressing need for scalable diagnostic tools that are both non-invasive and capable of timely and accurate detection of cognitive decline. Current neuropsychological assessments require clinical supervision, making them impractical for continuous or large-scale screening. Spontaneous speech, which captures natural, unscripted language use, offers a scalable alternative. By leveraging machine learning techniques, speech data can be analyzed to detect early signs of cognitive impairment.

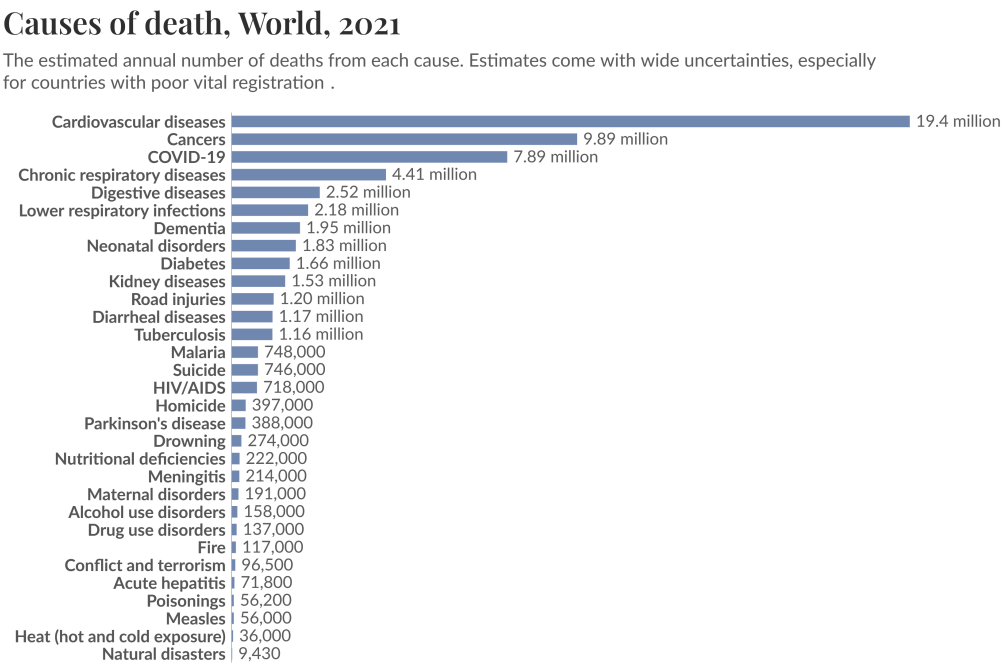

Figure 1. Annual number of deaths by cause. Source: OurWorldInData.org.

Previous studies [4], such as the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge [5], have shown that speech analysis can be used to detect cognitive decline. However, the trade-off between model performance and interpretability remains unresolved. Advanced deep learning models often achieve higher accuracy but function as "black boxes," limiting their clinical utility. This research addresses this challenge by primarily focusing on traditional machine learning models, which are trained on a combination of interpretable features and deep learning-derived embeddings. The study explores how these features can be used in conjunction with traditional models to detect cognitive decline, with a focus on balancing predictive accuracy and interpretability to ensure clinical relevance.

1.2 Background

Cognitive decline, encompassing conditions such as mild cognitive impairment and dementia, poses significant challenges for healthcare systems globally. Detection is crucial for enabling timely interventions that can delay the progression of these conditions and improve the quality of life for affected individuals. However, conventional diagnostic methods, such as MoCA and MMSE, are time-intensive and require specialized clinical expertise, making them unsuitable for large-scale or continuous screening. This limitation has driven the search for scalable, non-invasive tools that can complement traditional diagnostic approaches.

One promising avenue is the analysis of spontaneous speech, which reflects natural, unscripted language use. Speech is a complex behavior influenced by cognitive and neural processes, and changes in speech patterns often emerge in the early stages of cognitive decline. Common markers include reduced fluency, lexical retrieval difficulties, and incoherent speech organization. These markers align with the semantic, syntactic, acoustic, and informational impairments observed in conditions like Alzheimer's dementia and other dementias. By leveraging speech data, researchers can develop tools capable of detecting cognitive impairment more efficiently than conventional methods.

ML techniques have been increasingly applied to this domain, offering automated approaches to extract meaningful insights from speech data. Traditional ML models often utilize pre-defined linguistic and acoustic features, such as lexical diversity, pause frequency, and speaking rate. These features are interpretable and can be mapped directly to clinically observable speech patterns, making them useful for understanding a model's predictions. However, traditional ML models can also incorporate features derived from more advanced processing pipelines, including those output by deep learning models. In such cases, while the model itself may remain interpretable, the features it depends on may not be easily understood, thus introducing a layer of complexity in feature interpretability.

More complex neural models can bypass manual feature engineering by learning feature representations directly from raw data. This capability allows these models to capture the richness and subtlety of spontaneous speech, achieving higher accuracy in complex classification tasks. However, their complexity poses significant challenges for interpretability, as the learned features and the model's decision-making processes often lack transparency. This distinction highlights the dual challenge of interpretability: understanding both the features used by a model and the internal workings of the model itself.

Recent efforts, such as the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge [5] and its successors, have established benchmarks for evaluating speech-based cognitive assessment methods. These initiatives have advanced modeling approaches and led to the development of more robust datasets with reduced biases and confounding factors, highlighting the increasing potential of speech analysis to complement traditional diagnostic methods. By automating the prediction of cognitive assessment scores, there is significant potential to enhance the accuracy, reliability, and scalability of cognitive decline detection. This study builds on these advancements by evaluating different model–feature set combinations, with a focus on their clinical applicability and potential to improve speech-based diagnostic tools.

1.3 Related Work

Recent advances in speech processing have emphasized the potential of spontaneous speech as a non-invasive diagnostic tool for Alzheimer's dementia detection. Research in this area focuses on data collection, feature extraction, and classification methods, with growing interest in leveraging linguistic and acoustic markers of cognitive decline. Several comprehensive reviews offer an overview of this area, identifying key challenges, trends, and gaps. Voleti et al. [6] create a taxonomy for speech and language features indicative of cognitive and thought disorders including a variety of neurological impairments and psychiatric conditions. The work of de la Fuente Garcia et al. [4] targets cognitive decline specifically in the context of Alzheimer's disease, while Qi et al. [7] provide a more up-to-date analysis of the state-of the-art in AD detection.

Speech-based cognitive impairment research typically uses two main feature types: text-based and acoustic. Text-based features, such as type-token ratio and idea density, measure lexical and syntactical complexity, while indices like the Yngve and Frazier scores assess working memory. Acoustic features focus on prosodic elements like pause patterns and spectral features such as Mel-Frequency Cepstral Coefficients (MFCCs). In AD, these features reveal changes in speech due to cognitive and physiological impacts, such as imprecise articulation, altered vocal quality, and slower speech with more pauses. For Alzheimer's dementia detection, relevant acoustic features span frame-level descriptors (e.g., MFCCs, F0), deep learned embeddings (e.g., VGGish, Wav2vec 2.0), and higher-level prosodic measures. AD causes reduced fluency, word retrieval issues, and changes in sentence structure, which affect speech rhythm and acoustic energy distribution. Other indicators include dysfluency, linguistic features like vocabulary richness, and emotional and meta-features like age and gender. Combining handcrafted features with deep learning models is becoming more common for accurate AD detection.

As noted in [4] the vast majority of studies in AD detection utilize binary classification models to differentiate speech from Alzheimer's patients and healthy controls using acoustic and linguistic features. Only a limited number explore distinctions between MCI and HC/AD [8] or tackle three-way classifications (e.g., HC, MCI, AD). Most early studies utilise the Pitt Corpus collected by the University of Pittsburgh and distributed through DementiaBank [9]. Recently it has been shown that models trained on silent portions of this dataset can achieve almost perfect classification due to the presence of the Clever Hans effect [10], indicating the untrustworthiness of the results in these early works. Denoised and normalized versions of this corpus used in AD detection challenges (Interspeech 2020 and 2021) [5, 11] showed a 20 percent reduction in detection accuracy.

Several studies have explored the effectiveness of traditional machine learning models for this binary classification task. Wang et al. [12] found that the Support Vector Machine (SVM) classifier, when combined with BERT and Roberta features, achieved the best performance across five models, including LDA, Gaussian Process, Multilayer Perceptron (MLP), and Extreme Gradient Boost (XGB). Shah et al. [13] demonstrated that an ensemble of acoustic-based and language-based models outperformed individual models, with SVM, Logistic Regression (LR), and majority vote classifiers achieving strong results. Weiner et al. [14] used LDA for classification and achieved a high accuracy of 85.7 percent. Hernández-Domínguez et al. [15] found that SVM and Random Forest (RF) classifiers were effective for distinguishing between HC and MCI, offering valuable insights into early MCI diagnosis. Additionally, Edwards et al. [16] explored multiscale features at the word and phoneme levels, achieving a maximum classification accuracy of 79.2 percent using five models, including LDA, KNN, DT, RF, and SVM.

Deep learning models have shown significant promise in the classification of cognitive impairments such as AD. Warnita et al. [17] utilized a gated Convolutional Neural Network (CNN) to detect AD from speech data, achieving an accuracy of 73.6 percent. Other studies, like Koo et al. [18], explored improved convolutional Recurrent Neural Networks (RNNs), while Pan et al. [19] employed a bidirectional hierarchical RNN with an attention layer for AD detection. Ablimit et al. [20] combined CNN, bidirectional GRUs, self-attention, and Fully Connected Neural Networks for model fusion. Yang et al. [21] also applied a convolutional layer followed by LSTM layers for AD detection. Transformer-based models like BERT have also been effectively fine-tuned for AD detection, as seen in the work by Balagopalan et al. [22], where BERT was applied to speech samples. In addition, hybrid models combining deep learning techniques, such as CNN and RNN, along with speech features like MFCC, F0 envelope, jitter, and shimmer, have been used for dementia detection, with GCNN models reaching accuracies up to 73.6 percent [23]. The scarcity of publicly available data currently prevents large models from showing significant improvements over feature extraction and classification pipelines.

Only a limited number of studies address multi-class classification for distinguishing between HC, MCI, and AD. Mirzaei et al. [24] achieved 62 percent accuracy using a balanced French dataset of 48 participants, analyzing temporal and acoustic voice features (e.g., jitter, harmonics-to-noise ratio) from read speech with KNN, SVM, and Decision Tree (DT) classifiers. Similarly, Egas López et al. [25] achieved 56 percent accuracy with i-vectors extracted from MFCCs using an SVM classifier on a balanced Hungarian dataset of 75 participants. Gosztolya et al. [26] reported 66.7 percent accuracy using late fusion of acoustic and linguistic features from spontaneous speech, again with an SVM classifier, on a balanced dataset. In contrast, Kato et al. [27] achieved a higher

accuracy of 85.4 percent using an imbalanced Japanese dataset of 48 participants. Their study incorporated speech-prosody and cerebral blood flow activation during cognitive tests with a Bayesian approach. While these studies employ cross-validation techniques, they lack performance evaluation on a held-out test set, and none provide details on dataset accessibility. Furthermore, the datasets used in these studies are notably small, limiting the generalizability of the findings and emphasizing the need for larger, publicly available datasets for robust multi-class classification.

Studies on MMSE score prediction are relatively rare. One such study [11], using data from HC and AD individuals, employed an SVM regressor with eGeMAPS [28] acoustic features (e.g., F0 semitone, loudness, MFCCs, jitter, shimmer), second-order features from the active data representation method, and linguistic features from CHAT-compatible transcripts, achieving an RMSE of 5.29. Using a multilingual dataset that included MCI and HC subjects, [8] achieved an RMSE of 2.89 with a Multilayer Perceptron (MLP) model, leveraging linguistic features such as number of tokens, number of types, type-to-token ratio, density, verb ratio, and pronoun ratio extracted from ASR transcriptions.

Martin et al. [29] conducted a review on the state-of-the-art in interpretable machine learning for dementia diagnosis. Their work considers mainly imaging-based machine learning methods, with model interpretability as a specific inclusion criterion. They discussed the challenges in building robust, generalizable models and the growing importance of explainability in clinical applications. They highlight the promising classification performance of current models but also note the variability in validation procedures and the reliance on popular datasets and emphasize the need for clinician involvement in validating explanation methods and the critical analysis of interpretability techniques to ensure their applicability in clinical practice. While imaging-based approaches are noninvasive, they require specialized equipment and are more resource-intensive, making data collection more challenging than for speech-based methods.

Tan et al. [30] developed a machine learning model for early cognitive impairment diagnosis in a multi-ethnic Asian population, integrating socio-demographics, vascular risk factors, and neuroimaging markers. They applied SHapley Additive exPlanation (SHAP) [31] to identify key predictors, including age, ethnicity, education level, and neuroimaging markers, demonstrating the value of SHAP for model interpretability and reliable dementia diagnosis.

Iqbal et al. [32] applied Explainable AI techniques to speech-based Alzheimer's dementia screening, using linguistic features from speech transcripts of the Cookie Theft Picture Task. By employing Local Interpretable Model-agnostic Explanations (LIME) and SHAP, the

study identified key features influencing the model's decision-making. The model achieved 80 percent accuracy while providing transparent, interpretable results, which are important for clinical decision-making in AD diagnosis. However, the study's reliance solely on linguistic features limits its performance, and integrating acoustic and demographic data could improve both accuracy and clinical applicability.

Most research in AD detection has primarily focused on classification, particularly distinguishing between late-stage AD and healthy controls, with relatively little attention given to the detection of MCI and the prediction of the MMSE score, which are crucial for initial screening and early intervention. Although multi-modal approaches that combine acoustic, text-based, and demographic features show promise, relatively few studies explore these methods. The field also faces challenges in determining the true state of the art due to biased datasets, inconsistencies in data preparation and reporting, limited data availability, and variability in validation techniques, with cross-validation commonly used but held-out test sets less frequently applied. Additionally, while there has been significant work on the interpretability of imaging-based models and some exploration of linguistic features, there is a notable lack of studies on the interpretability of speech-based models for cognitive decline detection, which is important for their acceptance and effective use in clinical practice.

1.4 Research Objectives

The primary goal of this research is to evaluate and compare the performance and interpretability of different machine learning models for detecting cognitive decline through spontaneous speech. The study focuses on traditional machine learning models, which are primarily used to classify cognitive status and predict cognitive assessment scores. These models leverage both interpretable linguistic and acoustic features as well as deep learning-derived features, such as embeddings from fine-tuned deep learning models and features extracted from transcripts using large language models for classification and regression tasks.

Key research questions include:

1. How do machine learning models, using different combinations of both interpretable features and learned features from pretrained deep learning models, compare in classifying individuals as healthy, MCI, or dementia and predicting cognitive assessment scores?

2. Which acoustic, text-based, and demographic features—either automatically extracted or produced through deep learning—are most indicative of cognitive decline, and how do

they contribute to model decisions for both classification and regression tasks?

3. What is the trade-off between model performance and interpretability for the diagnosis and cognitive assessment score prediction tasks?

1.5 Structure of the Thesis

This thesis is structured as follows:

Chapter 1: Introduction. This chapter introduces the research topic, discussing the importance of cognitive decline detection and the potential of using spontaneous speech as a diagnostic tool. It presents the motivation for the study, highlights gaps in the existing literature, and defines the research objectives. Additionally, a brief overview of the related work is provided, outlining the current state of research in cognitive decline detection using speech. The chapter concludes with an outline of the structure of the thesis.

Chapter 2: Methodology. In this chapter, the methodology used to address the research objectives is outlined in detail. The data sources and preprocessing steps are described, along with the feature engineering process and the machine learning models used for detecting cognitive decline. The tasks and evaluation metrics employed to assess model performance are also discussed. Finally, the method for quantifying the interpretability of models, features, and their combinations is introduced.

Chapter 3: Results. This chapter presents the experimental results, beginning with an exploratory data analysis and followed by an evaluation of the models' performance on the cognitive decline detection tasks. A thorough analysis of the top performing features is provided, along with insights from the interpretability techniques applied.

Chapter 4: Discussion. This chapter summarizes the key findings of the research, outlines the limitations identified in the study, and discusses potential areas for future research.

Chapter 5: Summary. This chapter offers a concise summary of the thesis, revisiting the research objectives, methodology, and key findings, along with the main contributions of the work.

References. A comprehensive list of all sources cited throughout the thesis.

Appendices. The appendices include supplementary materials that support the findings of the study.

2. Methodology

This chapter outlines the methodology used to address the research objectives in detail. The data sources and preprocessing steps are described, along with the feature engineering process and the machine learning models used for detecting cognitive decline. The tasks and evaluation metrics employed to assess model performance are also discussed, followed by an explanation of the interpretability techniques used to quantify the interpretability of models, features, and their combinations.

The study employs a multi-step approach for detecting cognitive decline through spontaneous speech using machine learning models. The methodology involves a exploratory data analysis (EDA), feature engineering, a classification task, a regression task, and a focus on model interpretability, using a dataset from the Prediction and Recognition of Cognitive declinE through Spontaneous Speech (PROCESS) Signal Processing Grand Challenge, part of ICASSP 2025 [33].

Computations for this thesis were carried out using the high-performance computing infrastructure provided by TalTech [34]. The analysis and modeling were implemented primarily in Python, using libraries such as Scikit-learn [35], Pandas [36], NumPy [37], SHAP [38], XGBoost [39], and Plotly [40].

2.1 Data Sources and Preprocessing

The main dataset contains spontaneous speech recordings from individuals categorized as healthy, having mild cognitive impairment, or dementia. In addition to the main dataset, data from previous speech-based cognitive decline challenges (e.g., ADReSS [5], ADReSS-o [11], ADReSS-M [41], and TAUKADIAL [8]) is incorporated for augmentation. Compatibility between these external datasets and the PROCESS dataset is evaluated through statistical analyses, ensuring they augment the training data without introducing bias or inconsistency.

2.1.1 Main Dataset: PROCESS

The primary dataset used in this study is the PROCESS Signal Processing Grand Challenge dataset, which consists of speech recordings from individuals categorized as HC, MCI, or AD. The dataset includes recordings from three structured speech tasks:

- **Semantic fluency task (SFT):** Participants name as many animals as possible within one minute.
- **Phonemic fluency task (PFT):** Participants list words beginning with the letter P in one minute, excluding proper nouns.
- **Cookie Theft picture description task (CTD):** Participants describe a standardized picture (see Figure 2) from the Boston Diagnostic Aphasia Examination [42].

.

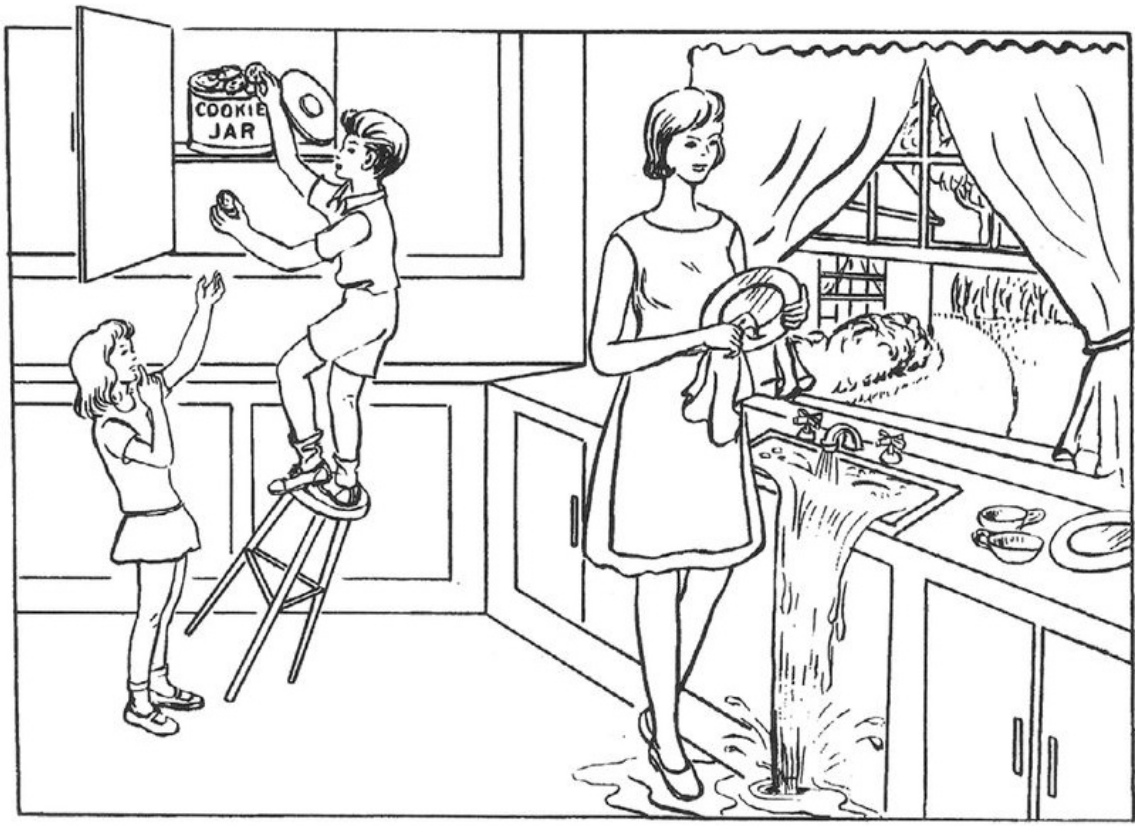

Figure 2. Cookie Theft picture from the Boston Diagnostic Aphasia Examination.

The dataset includes 157 participants, distributed as 82 HC, 59 MCI, and 16 AD, and contains manually transcribed speech samples. The dataset supports two tasks: classification of cognitive status (HC, MCI, AD) and regression predicting Mini-Mental State Examination (MMSE) scores. Performance evaluation is based on the macro-F1 score for classification and root mean squared error (RMSE) for regression.

2.1.2 External Datasets for Augmentation

To enhance model robustness and generalizability, additional datasets from previous speech-based cognitive decline challenges are incorporated. These datasets have been selected

with the assumption of compatibility with PROCESS, and statistical analyses are conducted to assess whether they introduce any bias or inconsistency.

ADReSS Dataset

The Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) dataset contains 156 recordings (78 AD, 78 non-AD) of participants describing the Cookie Theft picture. The dataset is gender- and age-balanced to minimize bias. The dataset includes manual transcripts annotated using the CHAT coding system and has undergone preprocessing steps such as noise removal and volume normalization.

ADReSS-o Dataset

The ADReSS-o dataset expands on ADReSS by including two distinct datasets:

1. Speech recordings from an AD cohort performing a semantic fluency task, used for predicting cognitive decline over two years.
2. Cookie Theft picture descriptions from both cognitively normal individuals and AD patients.

The dataset comprises 237 recordings split into training and test sets with a 70/30 ratio. Gender and age balancing was ensured through a propensity score matching approach.

ADReSS-M Dataset

The ADReSS-M dataset includes spontaneous speech descriptions of the Cookie Theft picture in English for training and a different picture in Greek for testing. The dataset is designed to study cross-linguistic cognitive assessment. The English training dataset is balanced for age and gender using propensity score matching. The Greek test set includes a task evaluating verbal fluency and mood using a standardized picture description protocol.

TAUKADIAL Dataset

The TAUKADIAL dataset contains English and Chinese speech samples from participants describing various pictures. The English dataset includes descriptions of the Cookie Theft, Cat Rescue, and Coming and Going images, while the Chinese dataset features three culturally relevant pictures. Participants were classified as MCI or normal cognition, and the dataset includes 507 total recordings (261 Chinese, 246 English), with an approximate 3:1 training-to-test ratio. Standardized propensity score matching was applied to ensure balanced age and gender distributions.

2.2 Feature Extraction

Three main types of features are extracted for the purposes of the classification and regression tasks: acoustic, text-based, and demographic features. Each of these have some variants that are interpretable and others which are derived from deep-learning models. An overview of the feature extraction pipeline is provided in Figure 3. Loudness normalization was performed on all audio files before any transcription, feature extraction, or downstream processing.

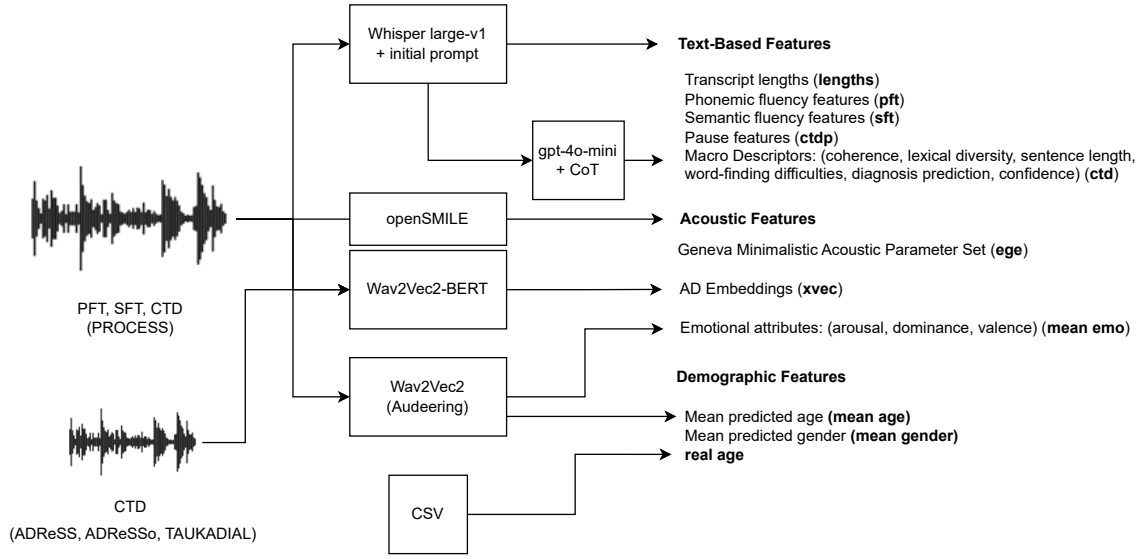

Figure 3. Feature extraction pipeline.

2.2.1 Acoustic Features

eGeMAPSv02 Low-Level Descriptors (`ege`): The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPSv02) is a widely used feature set for speech analysis, designed to capture key acoustic markers related to prosody, voice quality, and articulation. These features are defined as low-level descriptors and include several spectral, temporal, and energy-related measures. In this work, we extract eGeMAPSv02 features at 100ms intervals using openSMILE [43] and compute their mean and standard deviation over the entire recording. This approach captures both the central tendency and variability of each feature, effectively integrating temporal dynamics into the analysis. In the openSMILE implementation, depending on the feature, smoothing is applied either across all frames (_sma3) or only across non-zero frames (_sma3nz), with the latter helping to avoid distortions caused by unvoiced segments or missing values. The extracted low-level descriptors from eGeMAPSv02 can be categorized into the following groups:

1. Energy / Amplitude Features: these features reflect vocal intensity and voice quality, which can be indicative of vocal strength and control.

- **Loudness** (`Loudness_sma3`): perceived intensity of the speech signal.
- **Harmonics-to-Noise Ratio** (`HNRdBACF_sma3nz`): ratio of harmonic energy to noise energy, indicating voice clarity.
- **Shimmer** (`shimmerLocaldB_sma3nz`): cycle-to-cycle variation in amplitude, linked to voice stability.

2. Frequency Features: these features capture pitch characteristics and formant structure, which can be affected by neurological and cognitive decline.

- **Fundamental Frequency (F0)** (`F0semitoneFrom27.5Hz_sma3nz`): perceived pitch of speech, extracted in semitone scale.
- **Jitter** (`jitterLocal_sma3nz`): short-term irregularity in pitch, associated with unstable phonation.
- **Formants (F1, F2, F3) & Bandwidths** (`F1frequency_sma3nz`, `F1bandwidth_sma3nz`, `F2frequency_sma3nz`, `F2bandwidth_sma3nz`, `F3frequency_sma3nz`, `F3bandwidth_sma3nz`): resonant frequencies of the vocal tract, influencing articulation and vowel clarity.

3. Spectral Balance Features: These features describe the distribution of spectral energy and its evolution over time, relevant to articulation and phonation control.

- **Alpha Ratio** (`alphaRatio_sma3`): ratio of energy above and below 1 kHz, related to spectral tilt.
- **Hammarberg Index** (`hammarbergIndex_sma3`): ratio of strongest peak below and above 2 kHz, linked to spectral dominance.
- **Spectral Slope (0–500 Hz, 500–1500 Hz)** (`slope0-500_sma3`, `slope500-1500_sma3`): rate of energy decay in different frequency bands.
- **Mel-Frequency Cepstral Coefficients (MFCC 1–4)** (`mfcc1_sma3`, `mfcc2_sma3`, `mfcc3_sma3`, `mfcc4_sma3`): represent spectral shape and filterbank energies.
- **H1-H2 & H1-A3** (`logRelF0-H1-H2`, `logRelF0-H1-A3`): harmonic energy differences, indicating voice quality and phonation type.
- **Spectral Flux** (`spectralFlux_sma3`): measure of spectral change over time, indicative of speech dynamics.

Emotional Attributes (`mean_emo`): Emotional attributes capture key aspects of an individual's emotional state, which can be crucial for understanding psychological or affective changes in speech. For this work, arousal, dominance, and valence are extracted using a pre-trained wav2vec2 model, specifically fine-tuned on emotional speech data for dimensional emotion recognition. This model[1] [44], allows for the extraction of the following emotional dimensions:

- **Arousal** (`mean_emo_arousal`): this dimension represents the level of excitement or intensity in speech, ranging from calm to agitated. Higher values indicate more intense emotional states, while lower values reflect calmer emotional expressions.
- **Dominance** (`mean_emo_dominance`): this dimension reflects the control or power conveyed in speech, ranging from submissive to dominant. Higher values suggest greater perceived dominance, while lower values indicate more submissive tones.
- **Valence** (`mean_emo_valence`): this dimension represents the emotional valence, or the positivity/negativity of the speech. Higher values indicate positive emotions (e.g., happiness), while lower values are associated with negative emotions (e.g., sadness or anger).

These emotional attributes provide important insights into the affective state of the speaker, which could be indicative of psychological or cognitive states relevant to the study of cognitive decline.

AD Embeddings (`xvec`): For detecting cognitive decline directly from speech, we use a multilingual wav2Vec2-BERT model[2], which has been fine-tuned for Alzheimer's dementia prediction. This model was trained by aggregating the outputs of the wav2Vec2-BERT model with a statistics pooling layer, followed by a fully connected layer with ReLU activation and BatchNorm, ultimately producing a final output layer with three classes: HC, MCI, and AD. The model was trained using cross-entropy loss on randomly selected 5- to 8-second chunks of speech from AD-labeled utterances derived from the ADReSS, ADReSSo, and TAUKADIAL datasets. LoRA was employed to fine-tune the pre-trained model, with a configuration ($\texttt{rank} = 32$, $\alpha = 32$ and $\texttt{dropout} = 0.05$)

To extract high-level features, the trained AD classification model was repurposed as an embedding extractor. Specifically, 512-dimensional utterance embeddings were obtained from the output of the first dense layer (after the pooling layer), using the first 30 seconds

[1] https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim
[2] https://huggingface.co/facebook/w2v-bert-2.0

of each speech recording. These embeddings were used as input features for downstream tasks, with the aim of capturing speech patterns indicative of cognitive impairment. While not guaranteed to isolate disease-specific markers, such embeddings can encode complex, high-dimensional characteristics that may be missed by traditional acoustic features.

2.2.2 Text-based Features

Each recording was transcribed using the Whisper model (large-v1) [45], which provided word-level timestamps. The model was prompted with an initial instruction to better handle disfluent speech, and during decoding, the beam size was set to 5. The time-stamped transcripts were converted into a format matching the hand-annotated transcripts of the original PROCESS training dataset, where pause durations were given in parentheses.

Transcript Lengths (`lengths`): The transcript length refers to the total number of words spoken during the recording. This feature provides an indication of the verbosity of a speaker's response and is used as a simple measure of linguistic productivity.

Phonemic Fluency Features (`pft`): Phonemic fluency refers to the ability to generate words beginning with a specific letter or sound within a fixed period of time. In this study, phonemic fluency is assessed by counting the number of words starting with the letter P as well as other letters, along with several pause-related metrics. Specifically, we extract the following phonemic fluency features:

- **PFT count** (`p_words_pft`): the number of words starting with the letter P.
- **Other letters count** (`non_p_words_pft`): the number of words starting with letters other than P.
- **Pause count** (`count_pauses_pft`): the number of pauses during the phonemic fluency task.
- **Pause average length** (`avg_pause_pft`): the average length of pauses during the task.
- **Pause longest** (`longest_pause_pft`): the longest pause during the task.
- **Pause total duration** (`total_pause_pft`): the total duration of all pauses during the task.

These features allow us to analyze the speaker's ability to quickly produce words, while also identifying speech disruptions, which can be indicative of cognitive decline.

Semantic Fluency Features (`sft`): Semantic fluency refers to the ability to generate words from a specific category, such as animals, in a fixed amount of time. It is a key

measure for assessing cognitive function, as it reflects the ability to retrieve and produce semantically related concepts. In this study, we assess semantic fluency by extracting the following features:

- **SFT count** (`animals_sft`): the number of animals named during the task.
- **Repetitions count** (`repeats_sft`): the number of repeated words during the task.
- **Pause count** (`count_pauses_sft`): the number of pauses during the semantic fluency task.
- **Pause average length** (`avg_pause_sft`): the average length of pauses during the task.
- **Pause longest** (`longest_pause_sft`): the longest pause during the task.
- **Pause total duration** (`total_pause_sft`): the total duration of all pauses during the task.

These features are indicative of a speaker's ability to quickly retrieve words within a category.

CTD Pause Features (`ctdp`): Pause features measure the frequency and duration of pauses in speech, which can be indicative of cognitive load or difficulties in speech planning and execution. These features may provide insight into the speaker's cognitive processing, particularly in the context of Alzheimer's dementia and other forms of cognitive decline. The pause-related features extracted include:

- **Number of pauses** (`count_pauses_ctd`): the total count of pauses occurring in the speech.
- **Average pause length** (`avg_pause_ctd`): the mean length of all pauses.
- **Longest pause** (`longest_pause_ctd`): the longest single pause.
- **Total pause duration** (`total_pause_ctd`): the cumulative duration of all pauses in the speech.

These pause features help in understanding the pauses in relation to speech fluency and may highlight difficulties in speech production related to cognitive decline.

Macro-descriptors (`ctd`): Macro-descriptors are high-level features extracted from the speech transcripts that may offer insights into the overall quality of speech production. These features are designed to capture speech characteristics related to cognitive and linguistic function. The macro-descriptors include:

- **Coherence** (`coherence_pft`): measures the logical flow and connectivity of speech.
- **Lexical diversity** (`lexical_diversity_ctd`): assesses the variety of different words used in the speech, often linked to cognitive flexibility.
- **Sentence length** (`sentence_length_ctd`): the average number of words per sentence, reflecting sentence complexity and structure.
- **Word-finding difficulties** (`word_finding_difficulties_ctd`) : the frequency of hesitations, pauses or dysfluencies due to difficulty in retrieving words.
- **Alzheimer's prediction** (`alz_prediction_ctd`): a feature derived from the previous scores that estimates the likelihood of Alzheimer's disease based on language use.
- **Confidence in prediction** (`confidence_ctd`): the model's confidence in its Alzheimer's disease prediction based on the speech transcript.

These features were extracted using chain-of-thought (CoT) prompting with OpenAI GPT-4o-mini-2024-07-18, following the methodology described in [46].

2.2.3 Demographic Features

Demographic features are derived from the metadata csv files accompanying the datasets, where available. Potential issues with missing values are addressed during the exploratory data analysis phase (Section 3.1). These features can provide additional context for the analysis of speech patterns, as they are known to influence various speech characteristics.

Age of the speaker (`real_age`): The age of the speaker is an important demographic feature that may influence speech patterns, such as speech rate, pitch, and articulation. Age-related changes in cognitive abilities, such as those seen in aging, are often reflected in speech production. In this study, age is used as a continuous variable.

Gender of the speaker: The gender of the speaker is another key demographic feature. Gender differences in speech characteristics, such as fundamental frequency, articulation rate, and voice quality, have been well-documented. In this analysis, gender is treated as a categorical variable with two possible values: male or female.

2.3 Machine Learning Models

In this study, two types of machine learning models are employed: traditional machine learning models for classification and regression tasks[3] and deep learning models for feature extraction. The extracted deep learning features are integrated with interpretable features using early fusion.

2.3.1 Classification Models

Traditional machine learning models with text-based, acoustic, and demographic features are used to classify cognitive status and predict cognitive assessment scores. The following classification models are used in this study:

- **Random Forest (RF)**: An ensemble method that combines multiple decision trees to improve classification accuracy and reduce overfitting.
- **Logistic Regression (LR)**: A linear model used for binary and multiclass classification tasks. Logistic Regression is known for its simplicity, interpretability, and efficiency in handling linear decision boundaries.
- **Decision Tree (DT)**: A tree-based model that splits data into subgroups based on feature values. While easy to interpret, Decision Trees are prone to overfitting without proper tuning.
- **XGBoost (XGB)**: A highly optimized implementation of gradient boosting, typically using decision trees as weak learners. XGBoost is known for its ability to handle large and complex datasets while providing strong predictive performance.
- **Support Vector Machine (SVM)**: A supervised model that finds the optimal hyperplane to separate data into distinct classes. SVM is particularly effective for high-dimensional data and can handle non-linear decision boundaries using kernel tricks.
- **Naive Bayes (NB)**: A probabilistic classifier based on Bayes' theorem, assuming independence between features. While simple, Naive Bayes works well for text classification tasks and in scenarios with less complex relationships between features.
- **Gradient Boosting (GB)**: A boosting technique where each model corrects the errors of the previous one. Gradient Boosting is an effective method for classification tasks, especially when dealing with complex and noisy data.
- **K-Nearest Neighbors (KNN)**: A non-parametric classifier that assigns labels based on the majority class among the k-nearest neighbors in the feature space. KNN is intuitive and simple but can become computationally expensive with large datasets.

[3]Many models have both classifier and regressor variants (e.g., XGBoost). Only the base abbreviation (e.g., XGB) is listed; regressor variants use an appended "R" (e.g., XGBR) and appear as needed.

- **Multilayer Perceptron (MLP) Neural Network**: A feedforward artificial neural network composed of multiple layers. MLP is capable of learning non-linear relationships in the data and is useful for capturing complex patterns in feature spaces.
- **Voting Classifier (VC)**: A model that combines multiple classifiers (e.g., RF, SVM, and LR) to make final predictions based on their weighted outputs. Voting Classifiers are designed to improve model robustness and classification accuracy through ensemble learning.

2.3.2 Regression Models

In addition to classification, several traditional machine learning models are applied to regression tasks, predicting continuous cognitive assessment scores. The models used for regression are similar to those in classification, with the following specific models:

- **Random Forest Regressor (RFR)**: An ensemble of decision trees that predict a continuous value by averaging the outputs of multiple trees.
- **Linear Regression (LIR)**: A simple model that fits a linear relationship between features and the predicted target variable. Linear regression is easy to interpret and serves as a baseline for more complex models.
- **Ridge Regression (RR)**: A regularized version of linear regression that introduces an L2 penalty term (the square of the magnitude of the coefficients) to control for multicollinearity and prevent overfitting. Ridge Regression helps improve generalizability in the presence of highly correlated features.
- **Lasso Regression (Lasso)**: Similar to Ridge, Lasso Regression applies a penalty term, but it uses L1 regularization, which can shrink coefficients to zero and effectively perform feature selection.
- **XGBRegressor (XGBR)**: The regression variant of XGBoost, which implements gradient boosting for continuous target variables. XGBRegressor is designed to provide high performance, especially on large datasets.
- **Decision Tree Regressor (DTR)**: A non-linear model that splits data based on feature values to predict a continuous target. While interpretable, prone to overfitting without tuning.
- **Support Vector Regressor (SVR)**: A variant of SVM used for regression tasks, SVR attempts to find a function that approximates the data within a specified error margin, while maximizing the margin for prediction.
- **K-Nearest Neighbors Regressor (KNNR)**: A non-parametric model that predicts the target value based on the average value of its k-nearest neighbors. KNNR is intuitive and simple but can be computationally expensive when dealing with large

datasets.

- **Multilayer Perceptron Regressor (MLPR)**: A neural network used for regression tasks, capable of learning non-linear relationships between input features and the target variable. MLPR can capture complex patterns in data but requires careful tuning to avoid overfitting.
- **Gradient Boosting Regressor (GBR)**: A regression model that uses boosting to correct the errors of previous models in the sequence. GBR is widely used for regression tasks and performs well on noisy data.

2.3.3 Feature Extraction Models

In addition to traditional machine learning models, pretrained deep learning models are used to extract learned features from raw speech and transcripts. These features are used in conjunction with traditional models for both classification and regression tasks. The following pretrained deep learning models are employed:

- **Wav2Vec2-BERT**: a transformer-based model that learns speech representations directly from raw audio. This model is fine-tuned on the dataset to extract high-level features (AD embeddings) from speech data, which are then integrated into traditional machine learning models for classification and regression tasks.
- **Large Language Models (LLMs)**: pretrained language models are used to generate macro-descriptors from the transcribed speech, extracting high-level linguistic features that can reflect cognitive and linguistic function. These features capture aspects of speech such as coherence, lexical diversity, and word-finding difficulties, and they complement traditional, interpretable features in classification and regression tasks.

By combining the strengths of traditional machine learning models with deep learning-derived features, this research aims to leverage both the interpretability of traditional models and the performance of deep learning models in detecting cognitive decline and predicting cognitive assessment scores.

2.4 Tasks and Evaluation Metrics

In this study, subjects are classified into three categories: HC, MCI, or AD. To evaluate the classification performance, the macro-F1 score are used as the primary metric. The macro-F1 score is calculated by taking the F1 score of each class and averaging them. This metric is particularly useful for dealing with class imbalances, as it gives equal weight to each class, regardless of its frequency. The formula for the macro-F1 score is:

$$\text{F1}_{\text{macro}} = \frac{1}{C} \sum_{i=1}^{C} \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{2.1}$$

where C is the number of classes (in this case, 3), and Precision_i and Recall_i are the precision and recall for class i, respectively.

For the regression task, where the goal is to predict cognitive assessment scores (e.g., MMSE), the root mean squared error (RMSE) are used as the primary evaluation metric. RMSE is a common metric for regression tasks, as it penalizes large errors more significantly than smaller ones. It is calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_{\text{true},i} - y_{\text{pred},i})^2} \tag{2.2}$$

where N is the number of samples, $y_{\text{true},i}$ is the true value, and $y_{\text{pred},i}$ is the predicted value for the i-th sample.

2.5 Feature Set Combinations and Model Training

In this study, several feature sets are used for training machine learning models, including acoustic, text-based, and demographic features. These feature sets are combined in all possible combinations to assess their impact on model performance. Each feature set may contain one or more individual features, and the number of combinations depends on the total number of feature sets used.

2.5.1 Feature Set Combinations

Let $F_1, F_2, \ldots, F_n$ represent the different feature sets, where n is the total number of feature sets. The total number of possible feature set combinations is given by:

$$\text{Combinations} = 2^n - 1 \tag{2.3}$$

The subtraction of 1 accounts for the case where no features are selected. Each combination is then used to train all models in the set of traditional machine learning models.

2.5.2 Model Training Process

The models are trained using the following process:

1. **Feature Set Combinations**: All possible combinations of the feature sets are used for training. Each combination of feature sets is passed to the models for training, ensuring that every possible combination is tested.
2. **Cross-Validation**: For each feature set combination, the model is trained using 5-fold stratified cross-validation. Stratification ensures that each fold contains a representative proportion of each class. The cross-validation process is repeated for 9 different stratification seeds, ensuring robust evaluation and better generalization of the model given a relatively small dataset.
3. **Model Evaluation**: During each fold, the model is evaluated based on the macro-F1 score for classification tasks and RMSE for regression tasks. These metrics provide insight into the model's performance, accounting for class imbalance (in the case of classification) and prediction error (in the case of regression).
4. **Final Evaluation**: After training, the final model is evaluated on a held-out test set to assess its unbiased performance and to determine how well the model generalizes to new, unseen data.

2.6 Interpretability Techniques

An important aspect of this research involves comparing the interpretability of the various model–feature set combinations. While model-agnostic post hoc methods like SHAP [31] can quantify the contributions of individual features to model decisions both locally and globally, and inherently interpretable models like decision trees provide direct insight into decision-making, a suitable composite interpretability score will also be sought. This score will accommodate the differing features used by various models, with the goal of determining Pareto optimality in the trade-off between model performance and interpretability.

2.6.1 Joint Interpretability Index (JII)

The Joint Interpretability Index considers aspects of feature sets, machine learning models, and their interactions to assign a combined interpretability score to different model–feature set combinations.

$$\text{JII} = w_{\text{MI}} \cdot \text{MI} + w_{\text{FI}} \cdot \text{FI} + w_{\text{MFI}} \cdot \text{MFI} \tag{2.4}$$

The score is composed of three main parts: Model Interpretability (MI), Feature Set Interpretability (FI), and model–feature Set Interpretability (MFI). The weights $w_{\text{MI}}, w_{\text{FI}}, w_{\text{MFI}}$ are normalized and sum to 1, ensuring that each component contributes proportionally to the final Joint Interpretability Index (JII).

Model Interpretability (MI): a weighted numerical score based on qualitative algorithm properties (additivity, sparsity, linearity, smoothness, monotonicity, transparency), weighted by w_i, where the weights sum to 1. Each property is initially scored from 0 to 2, where 0 indicates that the model does not exhibit the property, 1 indicates that the model may exhibit it under certain conditions (e.g., specific hyperparameter values), and 2 indicates that the model exhibits the property with the hyperparameters used in training. The final score is normalized to fall within the range $[0, 1]$.

$$\text{MI} = \frac{1}{2} \sum_{i=1}^{n} w_i \cdot p_i \tag{2.5}$$

where:

- p_i is the score for interpretability property i (from 0 to 2),
- w_i is the weight assigned to the interpretability property i.

Feature Set Interpretability (FI): a weighted numerical score composed of three main subcomponents. The weights $w_{\text{FIC}}, w_{\text{FU}}, w_{\text{FL}}$ sum to 1.

$$\text{FI} = w_{\text{FIC}} \cdot \text{FIC} + w_{\text{FU}} \cdot \text{FU} + w_{\text{FL}} \cdot \text{FL} \tag{2.6}$$

- **Feature Set Correlation (FIC)** measures feature redundancy, weighted by w_{FIC}. This score has two weighted subcomponents: **Mean Absolute Correlation (MAC)** and **Mean Distance Correlation (MDC)**.

MAC captures linear associations between features and is computed as the mean of the absolute values of the off-diagonal entries of the Pearson correlation matrix:

$$\text{MAC} = \frac{1}{d(d-1)} \sum_{i=1}^{d} \sum_{\substack{j=1 \\ j \neq i}}^{d} |\rho_{i,j}| \tag{2.7}$$

where d is the number of features and $\rho_{i,j}$ is the Pearson correlation coefficient between feature i and feature j.

MDC captures both linear and nonlinear associations and is computed using distance correlation:

$$\text{MDC} = \frac{1}{d(d-1)} \sum_{i=1}^{d} \sum_{\substack{j=1 \\ j \neq i}}^{d} \text{dCor}(X_i, X_j) \tag{2.8}$$

where $\text{dCor}(X_i, X_j)$ denotes the distance correlation between feature vectors X_i and X_j.
The final FIC score is then computed as a weighted average of the two subtracted from 1:

$$\text{FIC} = 1 - (w_{\text{MAC}} \cdot \text{MAC} + (1 - w_{\text{MAC}}) \cdot \text{MDC}) \tag{2.9}$$

where $w_{\text{MAC}} \in [0, 1]$ controls the relative importance of MAC vs. MDC in the final redundancy score.

- **Feature Set Understandability (FU)**: based on domain knowledge, weighted by w_{FU}. Each feature is assigned a qualitative interpretability score $q_i \in [0, 1]$. The FU score is the average over all features in the feature set:

$$\text{FU} = \frac{1}{d} \sum_{i=1}^{d} q_i \tag{2.10}$$

  where:
  - d: number of features in the selected feature set
  - q_i: qualitative interpretability score of feature i, inherited from its feature group

- **Feature Set Length (FL)**: penalizes longer feature sets, as they are generally considered less interpretable. To avoid undefined logarithms and control the scaling behavior of the penalty, a value of 1 is added to both the feature set length and the maximum dimensionality before applying the logarithm. The logarithm is used to introduce a diminishing returns effect, where the penalty for adding more features

to the set becomes less severe as the set grows larger. This reflects the intuition that adding features to an already complex set has a less significant impact on interpretability compared to adding features to a simpler set.

$$\text{FL} = 1 - \frac{\log\left(\frac{f}{s} + 1\right)}{\log\left(\frac{d_{\text{max}}}{s} + 1\right)} \tag{2.11}$$

where:

- f is the feature set length,
- s is the scaling factor,
- d_{max} is the maximum dimensionality for a feature set combination.

model–feature Set Interpretability (MFI): based on the entropy of normalized feature importance magnitudes derived from SHAP values, reflecting the distribution of global feature importance.

The model–feature Set Interpretability (MFI) score is calculated as:

$$\text{MFI} = 1 - \frac{H(X)}{H_{\text{max}}} \tag{2.12}$$

where:

- $H(X) = -\sum_{i=1}^{n} p_i \log_2(p_i)$ is the entropy of the normalized SHAP values, quantifying the spread of feature importance across the feature set.
- $H_{\text{max}} = \log_2(n)$ is the maximum possible entropy for n features, representing a uniform distribution of importance.
- p_i is the normalized SHAP importance of the i-th feature, representing its global importance relative to other features. It is calculated as follows:
  - For classification models, the absolute SHAP value for each feature is calculated by summing the absolute values of its SHAP values across all classes. This absolute SHAP value is then normalized so that the sum of all feature importances is 1.
  - For regression models, the absolute SHAP value for each feature is calculated by summing the absolute values of its SHAP values. This absolute SHAP value is then normalized similarly to ensure the sum of feature importances is 1.

Lower entropy indicates that a smaller number of features are primarily driving the model's predictions, leading to higher interpretability. MFI is normalized to be between 0 and 1, where 1 represents maximum interpretability (all importance is concentrated on a single feature).

2.6.2 Rationale Behind JII Components

The Joint Interpretability Index (JII) is designed to quantify the interpretability of machine learning models based on both their internal characteristics and the features they utilize. Each component of the JII targets different aspects of interpretability, and their combined score reflects an overall assessment of the model's interpretability. Below is the rationale for each of these components:

- **Model Interpretability (MI):** the interpretability of a model refers to how easily a human can understand the relationship between its inputs and outputs. Certain structural and behavioral properties influence this interpretability by making the model's behavior more transparent and predictable. These key properties include:
  - *Additivity*: the model's prediction is formed by summing distinct contributions from individual features or feature groups. This structure enables clear attribution of how each input affects the output.
  - *Sparsity*: only a small number of input features meaningfully influence the model's predictions. Sparse models are easier to interpret because they focus attention on a limited set of relevant variables.
  - *Linearity*: the effect of each input feature on the prediction is proportional and consistent. Linear relationships are inherently simple, making it easier to reason about how changes in inputs affect outputs.
  - *Monotonicity*: a feature's influence on the prediction moves consistently in one direction—either always increasing or always decreasing the output. This aligns with domain expectations and makes model behavior more intuitive.
  - *Smoothness*: the model responds gradually to small changes in input. This prevents abrupt or unpredictable shifts in predictions, increasing trust and making the model's behavior feel more stable and continuous.
  - *Transparency*: the internal structure of the model—such as coefficients in linear regression or decision paths in a tree—is directly accessible and understandable. Transparent models allow users to trace and verify predictions without complex post-hoc analysis.

  Models exhibiting these properties are typically considered as more interpretable

because their decision-making processes can be more easily understood, explained, and trusted. The inclusion of the properties in this component is inspired by [47, 48].

- **Feature Set Interpretability (FI)**: the interpretability of a feature set is determined by its redundancy, understandability, and size. A feature set that is less redundant and more understandable is easier to interpret.
  - *Redundancy* arises when features are highly correlated. High correlations between features make it difficult to disentangle their individual contributions to the model, reducing interpretability. Therefore, feature sets with lower correlations between features are favored.
  - *Understandability* refers to the alignment of features with domain knowledge. Features that are easily understood in the context of the problem at hand improve interpretability.
  - *Feature set size* also plays a role: smaller feature sets are generally easier to interpret because they provide a simpler and more concise representation of the data.

  Therefore, feature sets that are compact, less redundant, and easier to interpret based on domain knowledge are scored higher.
- **Model–Feature Set Interpretability (MFI)**: the interaction between a model and its feature set contributes significantly to overall interpretability. The MFI score is based on the entropy of the SHAP feature importances, which quantifies how concentrated or distributed the feature contributions are.
  - When feature importances are concentrated (i.e., a few features dominate), the model is easier to interpret because it is clear which features drive the model's predictions.
  - When feature importances are uniformly distributed, the model becomes more complex to understand since no single feature stands out as influential, making the decision process harder to explain.

  Models with a small number of highly influential features are considered more interpretable than those where feature importances are spread across many features, contributing to the final MFI score.

By integrating these components, the Joint Interpretability Index provides a comprehensive and structured metric for evaluating model interpretability, which is used for identifying model–feature set combinations that lie on the Pareto front balancing predictive accuracy and interpretability across both tasks. The weights of the individual components can be adjusted to reflect the preferences of different stakeholders, making the JII adaptable to various interpretability needs.

2.7 Validation

Model performance is compared to the baseline results from the PROCESS challenge. Cross-validation techniques are employed on the PROCESS dataset, with final evaluations conducted on a held-out test set. Features derived from deep learning models are expected to offer higher accuracy than traditional hand-crafted features, their interpretability is carefully assessed to ensure transparency in their predictions, which is essential for clinical applications.

3. Results

This chapter presents the results of the experiments conducted, starting with EDA and moving on to the performance of the models on the cognitive decline detection tasks, followed by an analysis of the features that contributed most to the model's predictions along with insights from the interpretability techniques applied.

3.1 Data Diagnostics and Methodological Adjustments

An exploratory data analysis was undertaken to assess dataset characteristics, identify inconsistencies, and uncover potential biases. Based on these findings, methodological adjustments are proposed to enhance data integrity and model reliability.

3.1.1 PROCESS dataset

The primary dataset in its original form was composed of 157 folders, titled after the unique IDs of each of the subjects. Each folder includes three recordings per subject for each speech elicitation task, these are the CTD, SFT, and PFT tasks as described in Section 2.1.1. The total duration of the combined 471 recordings is 8 hours 25 minutes and 42 seconds, with a mean duration of 1 minute and 4 seconds per recording. Table 1 presents the results of statistical analysis of the audio file durations (in seconds) categorized by their suffix. As seen, CTD.wav files have the highest mean duration, while PFT.wav and SFT.wav have more consistent durations with lower standard deviation.

Table 1. Statistics of audio file durations by category.

	Count	Mean	Std Dev	Min	Max
CTD.wav	157.0	72.9276	39.0753	8.56	190.448
PFT.wav	157.0	59.9724	3.6360	41.67	71.830
SFT.wav	157.0	60.3613	3.0952	45.83	70.880

Each folder also includes three manually transcribed text files corresponding to each of the recordings per subject. The transcripts include the duration of pauses in seconds in parentheses.

In addition to the audio and text files, the dataset includes a CSV file including demographic information about each speaker. The CSV file includes 157 rows and 6 columns. The columns are:

- **Record-ID:** The unique ID of the speaker.
- **TrainOrDev:** Subset the recording belongs to (train/dev).
- **Class:** The diagnostic label of the speaker (HC/MCI/AD).
- **Gender:** The gender of the speaker (male/female/other).
- **Age:** The age of the speaker.
- **Converted-MMSE:** The MMSE score of the speaker.

An overview of the diagnostic class and gender distribution in the dataset is provided in Table 2. There are no missing values.

Table 2. PROCESS train/dev distribution by class and gender.

TrainOrDev	AD	HC	MCI	female	male	other	Count
train	12	61	44	63	53	1	117
dev	4	21	15	18	22	0	40
total	16	82	59	81	75	1	157

In the Age and Converted-MMSE columns, a high number of values are missing, approximately 56% and 20% respectively, see Table 3. The missing age values have been replaced by the mean age of the rest of the dataset, which is 66 years.

Table 3. Summary statistics for age and converted-MMSE.

	Age	Converted-MMSE
count	126.00	69.00
mean	65.72	27.36
std	13.74	2.47
min	23.00	19.00
25%	61.00	27.00
50%	69.00	28.00
75%	73.75	29.00
max	94.00	30.00
missing	31.00	88.00
missing percent	19.75%	56.05%

There is noticeable overlap between the missing values in the age and MMSE columns, with 33% of all missing MMSE values also missing the age value, see Figure 4). Among the subjects who are missing both age and MMSE scores, 7 are MCIs and 22 are HCs.

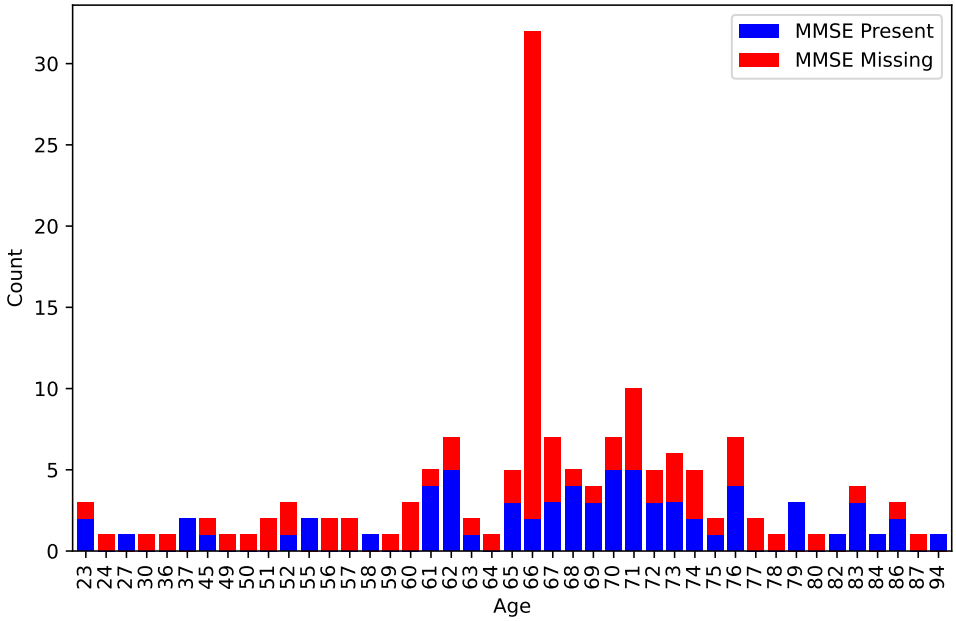

Figure 4. Age distribution: MMSE present vs MMSE missing.

Also of interest is the relationship between age and the dignostic class, see Figure 5. Included here are only data points where the age value is present. In general, subjects in the Dementia class tend to be older, and multiple outliers can be seen among the healthy controls.

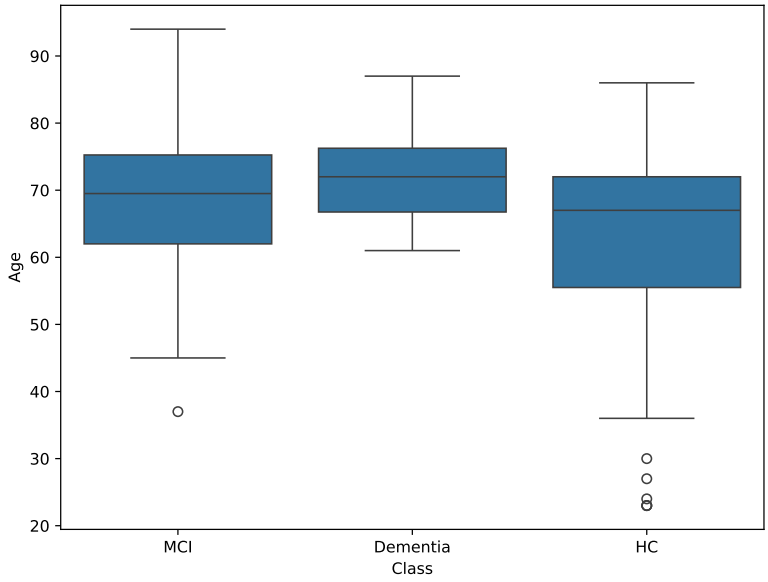

Figure 5. Age distribution by diagnostic class.

Looking at MMSE distribution vs. diagnostic class shows that MCI and Dementia are difficult to differentiate based only on the MMSE score, see Figure 6.

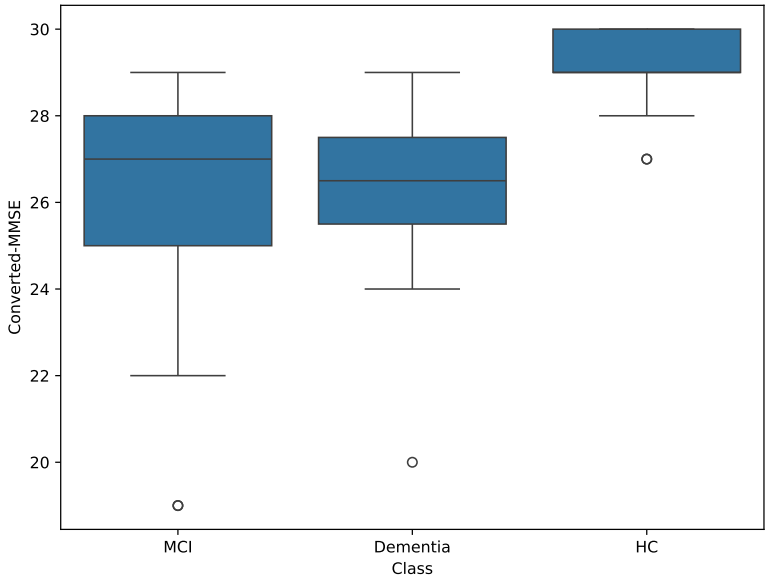

Figure 6. MMSE distribution by diagnostic class.

56% of missing MMSE values poses problems with an already small dataset. The 20% of missing ages are predominantly among the HC group, which is on average younger (61 years). Listening to samples from the HC group with missing age values suggests that they are on average younger than the group mean. This necessitated an alternative approach to filling the missing values, an extra feature called `mean_age` was added, which was predicted using a publicly available wav2vec-based model[1] [49]. To counter the possibility that the test set might not include gender data, a feature named `mean_gender` was produced using the same model. The model used to derive this feature is so consistent, it was able to identify two cases of mislabeled gender in the ground-truth data.

3.1.2 External data

The external datasets were all structured in a unique way, different from the PROCESS dataset. They include a variety of folders and files with metadata, which needed to be carefully combined to fully understand the contents of each dataset.

The ADReSS dataset includes separate folders for training and test sets. These folders hold three subfolders, one for transcriptions, one for full-wave enhanced audio, and one for

[1]https://huggingface.co/audeering/wav2vec2-large-robust-24-ft-age-gender

normalised audio chunks. In addition, there are two text files specifying the age, gender, and MMSE score for each subject.

The ADReSS-o dataset includes folders for two different tasks, progression and diagnosis. Each includes folders for the respective training and test sets. These folders hold two subfolders, one titled `segmentation`, which holds CSV files with speaker segmentation data for each recording, in subfolders for AD and HC patients, and another with subfolders for AD and HC that hold the audio files. In addition, there is a CSV file specifying the MMSE score and diagnostic label for each speaker. Age and gender information was present in separate CSV files and had to be combined with the previous ones.

The ADReSS-M dataset includes various folders and CSV files with ground-truth data. The train folder includes the training audio data. The `test-gr` folder holds the test set audio data, which is spoken in Greek language. There are also two folders, with combined 16 Greek samples. Uniquely, included in the metadata CSV for this challenge is the years of education for each speaker, in addition to gender and age.

The TAUKADIAL dataset includes two folders separating the training and test audio data. There are three different recordings for each speaker, all of which are related to picture description tasks, and the language of the audio is either English or Chinese. The metadata CSV file for the training and test sets includes information for the age, gender, MMSE score, and the diagnostic label of each speaker.

All datasets included the CTD recordings. Some of the external datasets were composed of a mixture of English and Chinese or Greek recordings. Various recordings were of different picture description task, and while some semantic fluency tasks, namely in ADReSS-o, did exist, these were not used for augmentation. From the 1356 full-length recordings an initial subset of 869 recordings were retained. Among these are 778 with an MMSE score, In addition to the 88 datapoints with missing MMSE scores in the PROCESS dataset, another 3 had missing MMSE score in the ADReSS, ADReSS-o, and ADReSS-M datasets, with just a combined 778 values with MMSE scores, however these missing values affect only the MMSE prediction task and not classification. An overview of the datasets considered for augmentation are presented in Table 4.

Table 4. Diagnosis and gender distribution by dataset.

	HC	MCI	AD	female	male	other	Count
PROCESS	82	59	16	81	75	1	157
ADReSS	78	-	78	86	70	-	156
ADReSS-o	115	-	122	154	83	-	237
ADReSS-M	115	-	122	154	83	-	237
TAUKADIAL	31	51	-	55	27	-	82

Looking at the age distribution by diagnostic class (Figure 7), it can be seen that for the healthy control group, the mean age is relatively homogeneous across all datasets, with PROCESS having many outliers on the younger side, and TAUKADIAL on the older side. These are also the only two datasets which contain MCI data, and in TAUKADIAL, the age is on average around 6 years higher. Also, the minimum age for the HC and MCI classes is noticeably lower in the PROCESS dataset compared to others.

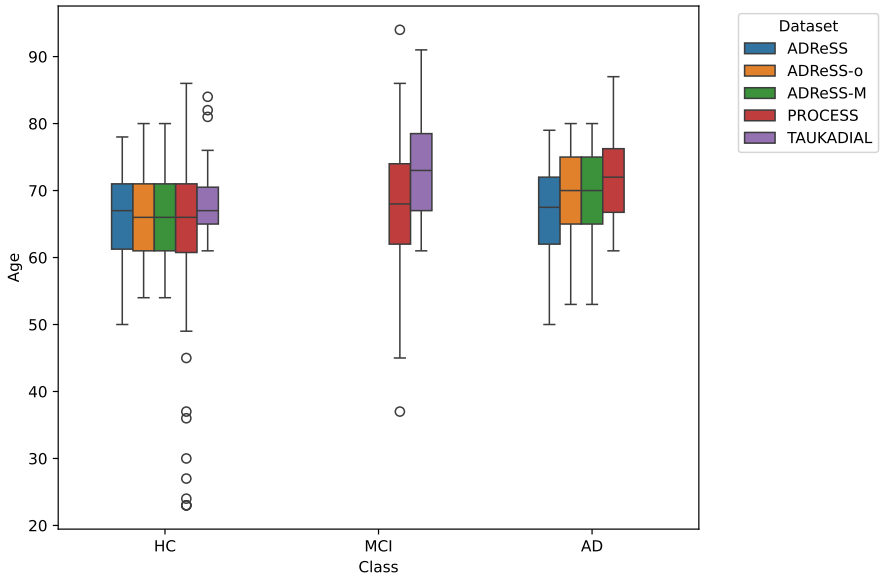

Figure 7. Age distribution by diagnostic class, combined dataset.

For additional information about the age statistics across the different diagnostic classes in each dataset see Table 5.

Table 5. Age statistics by dataset and diagnostic class.

Dataset	class	mean	std	min	max
PROCESS	AD	71.8	7.6	61	87
	HC	62.9	13.6	23	86
	MCI	68.2	10.3	37	94
ADReSS	AD	66.6	6.8	50	79
	HC	66.3	6.6	50	78
ADReSS-o	AD	69.4	6.9	53	80
	HC	66.1	6.3	54	80
ADReSS-M	AD	69.4	6.9	53	80
	HC	66.1	6.3	54	80
TAUKADIAL	HC	68.5	5.8	61	84
	MCI	73.5	7.6	61	91

Figure 8 presents an overview of the class-wise distribution of MMSE scores across the datasets. For TAUKADIAL and PROCESS datasets, the MMSE scores for HC class are on average higher. For the MCI class, PROCESS data includes subjects with slightly lower MMSE compared to TAUKADIAL. The most pronounced difference is among the AD class, where PROCESS has noticeably higher MMSE scores compared to others.

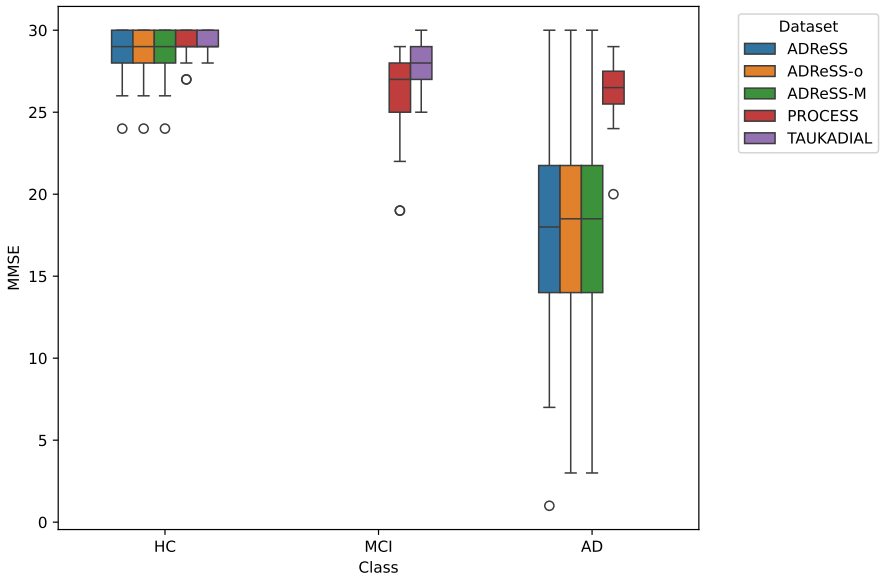

Figure 8. MMSE distribution by diagnostic class, combined dataset.

Gender proportions for each dataset are presented in Figure 9. In general, there are more female subjects, but in PROCESS the distribution is more balanced.

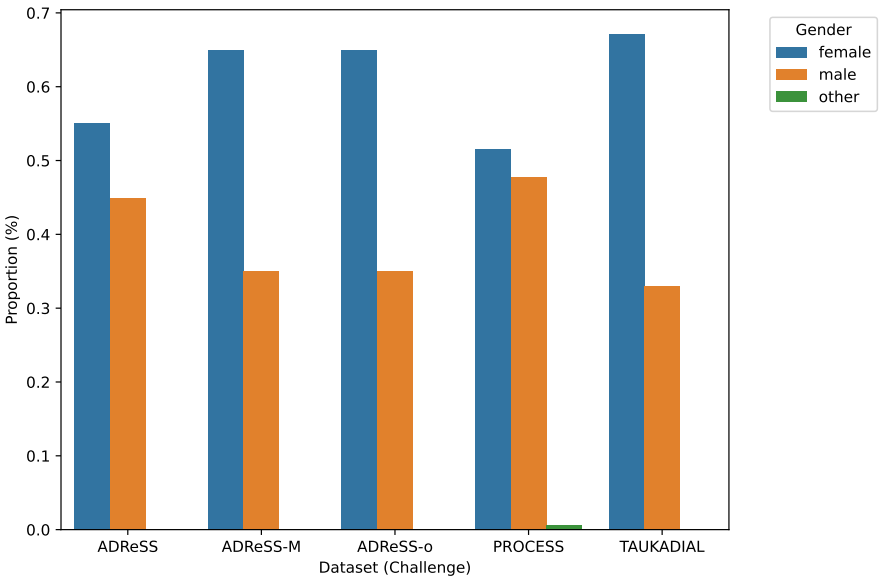

Figure 9. Gender proportions across datasets.

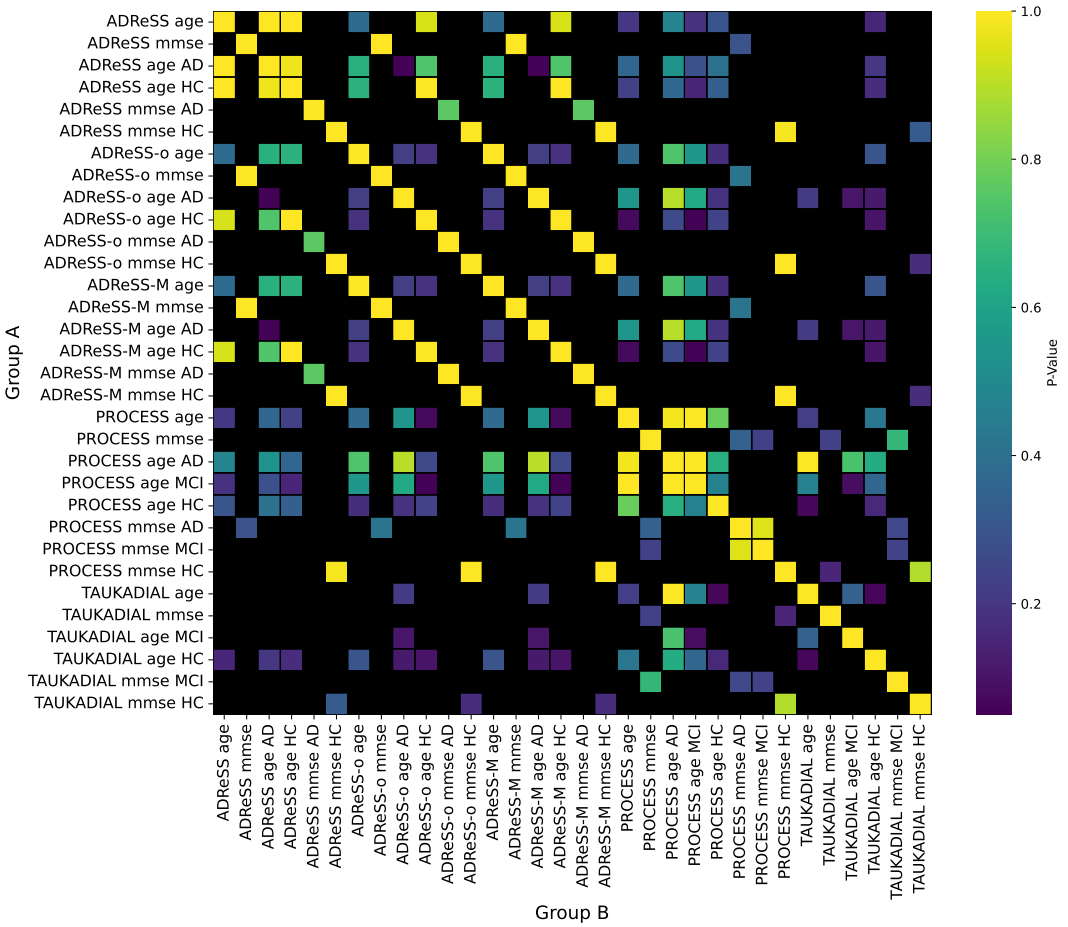

Figure 10. P-value matrix, Kolmogorov–Smirnov two-sample test.

Kolmogorov-Smirnov two sample test was used to determine if the distributions of class-

wise age and MMSE scores were similar across the datasets. Figure 10 presents the outcome, with black squares indicate a p-value with $p < 0.05$, meaning the difference is significant.

Many of the above figures strongly indicate that the ADReSS-o dataset and the English training set of ADReSS-M dataset are identical. The audio in ADReSS-M was of lower quality mp3 format, therefore it was dropped completely, and the recordings from ADReSS-o were kept. In addition, there was suspicion that a portion of the ADReSS dataset also overlapped with ADReSS-o. To investigate this, first, the durations and file sizes in both datasets were compared. Upon a match, the two files were listened to, and in case they were duplicates, the ADReSS version was removed from the dataset. This approach identified some additional 40 duplicates. However, this approach could not remove all the duplicates, as the quality of the recordings varied between the two datasets and at times slightly different duration versions of otherwise identical recordings were present.

To counter this, the next step in the exploratory data analysis was to extract text embeddings from all CTD recordings and use t-distributed Stochastic Neighbor Embedding (t-SNE) to visualize them in a lower dimensional space. One interesting finding revealed in Figure 11 was a group of outliers composed of MCIs to the far left of the plot.

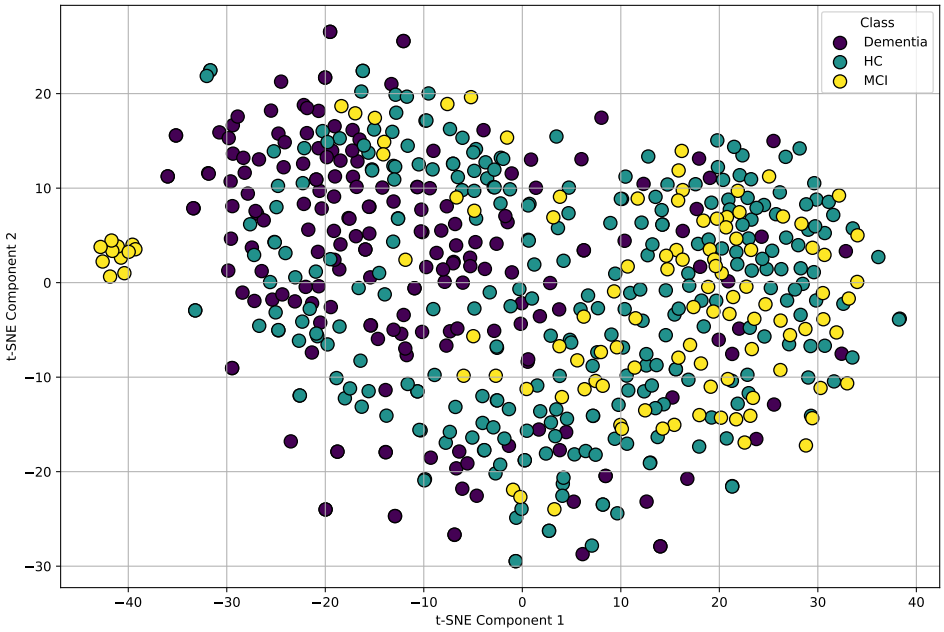

Figure 11. T-SNE of transcript embeddings by class (before).

The same data was visualized colored by dataset Figure 12. It appeared that the MCI group was in the TAUKADIAL dataset. Closer inspection revealed that the cluster of outliers were not embeddings of the transcripts of a CTD recording, but rather a different picture description task — a case of data discrepancy and mislabelling in the TAUKADIAL dataset. However, this figure revelead more problems, as it can be seen that in terms of t-SNE component 1, the majority of datapoints in the PROCESS dataset appear on the far right of the plot. With ADReSS and ADReSS-o on the left with some TAUKADIAL datapoints.

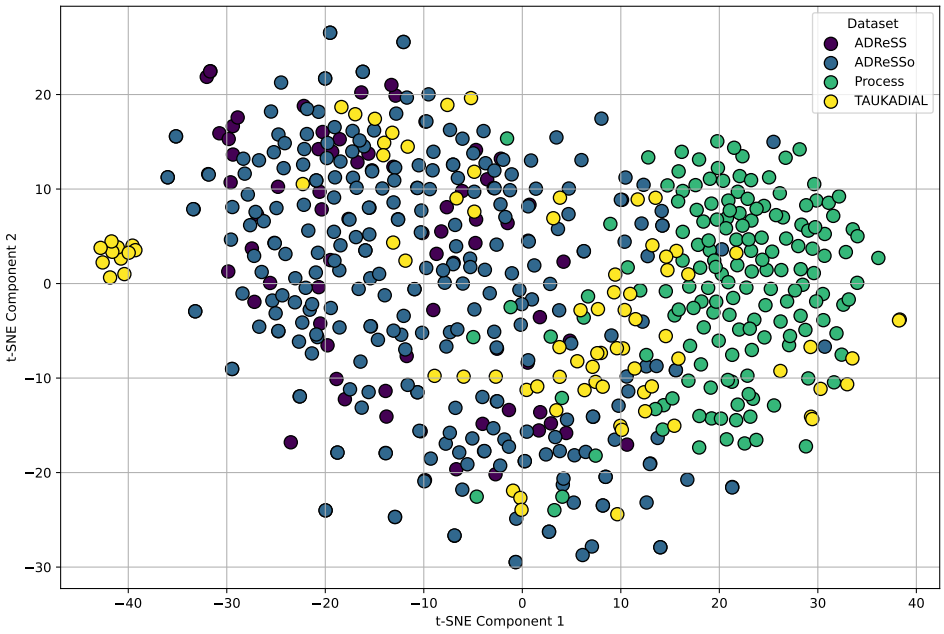

Figure 12. T-SNE of transcript embeddings by dataset (before).

An interactive version of the plot revealed the main factor that t-SNE component 1 seemed to encode: the presence of instructor's speech in the transcripts. In the TAUKADIAL and PROCESS datasets, and with minor exceptions, only the patient is speaking. While this may help distinguish datasets based on text embeddings, it introduces serious bias, as the majority of AD cases are in the ADReSS and ADReSS-o datasets. To combat this, speaker segmentation and diarization was undertaken with the aim to keep only the dominant speaker — the patient. While this approach worked relatively well for TAUKADIAL and PROCESS datasets, the extremely low quality of audio in the other two datasets resulted in only 10% of the diarization output to be accurate (it mostly still contained both speakers). As a result, the full combined dataset was recut manually. In addition, the interactive visualization revealed many additional overlapping datapoints, revealing further duplicates.

These were removed by comparing the similarity of a dimensionally reduced versions of text embeddings and eGeMAPS feature vectors for each pair of data points. The twofold approach was prompted by the fact that the output from the transcription model alone could not guarantee perfect matches. Figure 13 shows the text embeddings after removal of duplicates and keeping only patient speech. It can be seen that the magnitude of the t-SNE components has reduced a lot, and that the data are much more heterogeneously distributed.

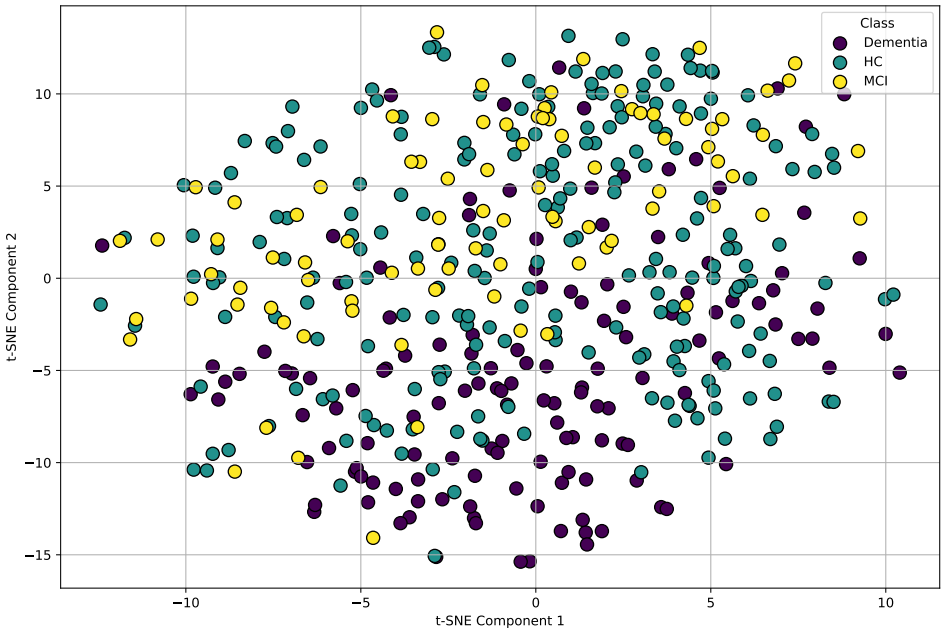

Figure 13. T-SNE of transcript embeddings by class (after).

Figure 14 shows the text embeddings by dataset. While some separation can be noticed, a clear decision boundary does not exist. In summary, while the text embeddings seem to distinguish somewhat between the different datasets, they do not show much usefulness in clearly separating HC, MCI, and AD individuals.

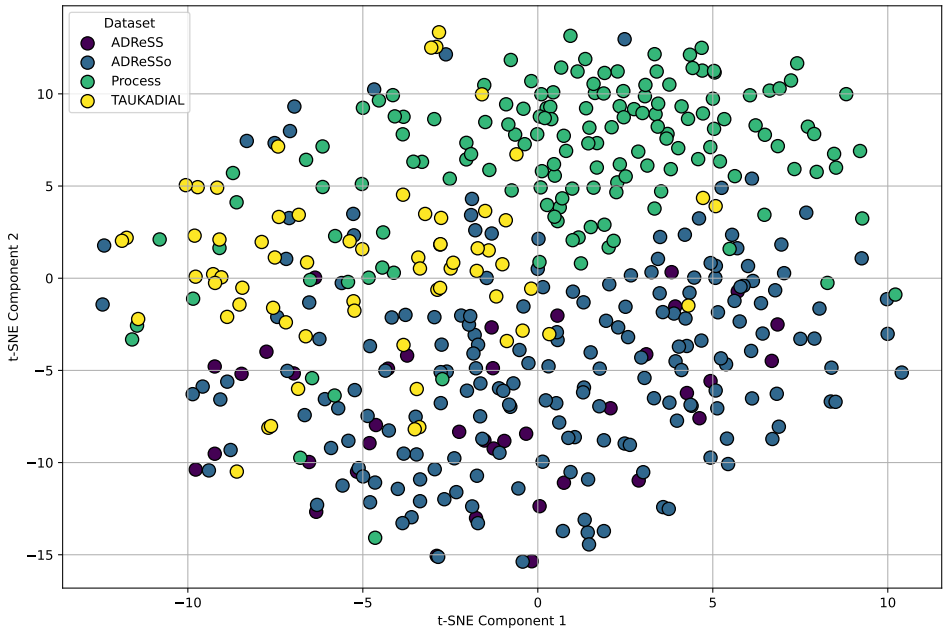

Figure 14. T-SNE of transcript embeddings by dataset (after).

The resulting combined dataset of total size 451 subjects, with the duplicates removed, is described in Table 6.

Table 6. Diagnosis and gender distribution by dataset (after).

	HC	MCI	AD	female	male	other	Count
PROCESS	82	55	16	79	73	1	153
ADReSS	17	-	21	21	17	-	38
ADReSS-o	92	-	101	128	65	-	193
TAUKADIAL	31	36	-	47	20	-	67

Figures 15 and 16 present the class-wise age and MMSE distributions of the final combined dataset.

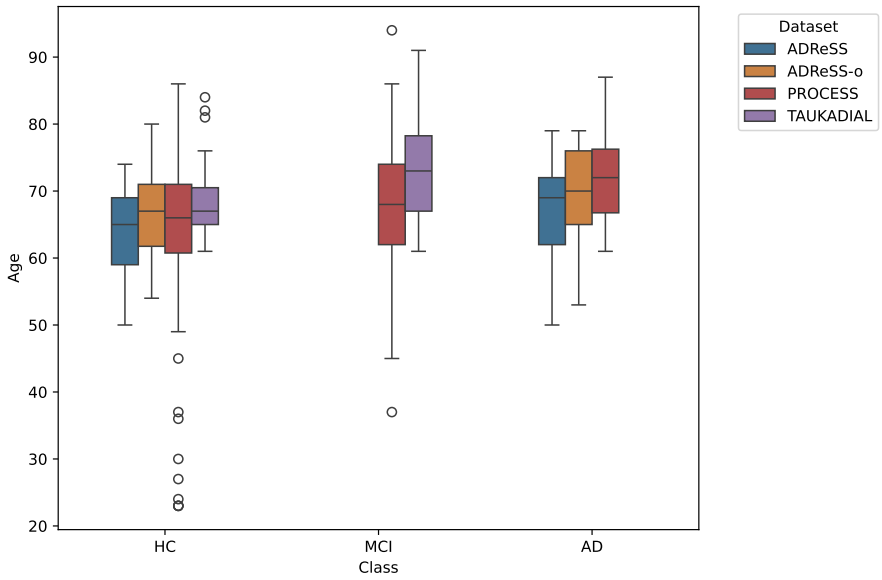

Figure 15. Age distribution by diagnostic class, combined dataset (after).

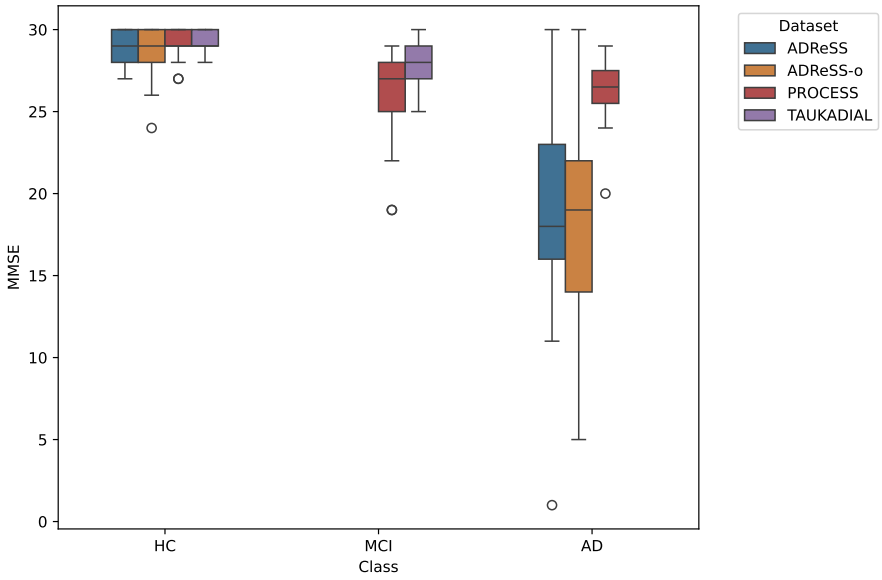

Figure 16. MMSE distribution by diagnostic class, combined dataset (after).

Additional statistical tests were applied to determine whether data from different challenges could be used for augmentation comparing the distributions of MMSE scores across datasets for each diagnostic class (HC, MCI, AD).

First the Kruskal-Wallis (KW) test was applied to assess whether there were significant differences in MMSE distributions across multiple challenges. If the KW test indicated a significant difference ($p < 0.05$), Mann-Whitney U (MW) tests were performed for pairwise comparisons to identify which datasets differed. See Table 7.

Table 7. Statistical test results for class-wise MMSE distributions across challenges.

Class	Test	Comparison	p-value
HC	KW	PROCESS vs ADReSS, ADReSSo, TAUKADIAL	0.9148
	MW	PROCESS vs ADReSS	0.7254
	MW	PROCESS vs ADReSSo	0.9102
	MW	PROCESS vs TAUKADIAL	0.5194
	MW	ADReSS vs ADReSSo	0.7444
	MW	ADReSS vs TAUKADIAL	0.9061
MCI	KW	PROCESS vs TAUKADIAL	**0.0433**
	MW	PROCESS vs TAUKADIAL	**0.0439**
AD	KW	PROCESS vs ADReSS, ADReSSo	**0.0006**
	MW	PROCESS vs ADReSS	**0.0028**
	MW	PROCESS vs ADReSSo	**0.0001**
	MW	ADReSS vs ADReSSo	0.9566

The results suggest that HC distributions do not differ significantly across datasets, meaning HC samples from other challenges could be considered for augmentation. However, MCI and AD distributions showed significant differences, particularly for AD in PROCESS vs. ADReSS/ADReSS-o, indicating that augmentation from these datasets may introduce inconsistencies. Given that HC is already the majority class in PROCESS, augmenting it further would not be beneficial, while augmentation for MCI and AD remains uncertain due to statistical differences. Using only the PROCESS dataset would yield $153 \times 3 = 459$ recordings — exceeding the combined dataset (451 CTD recordings) while still providing task variability and consistent audio quality. For this reason, the majority of the work henceforth was conducted using only the PROCESS dataset. The external data was used only for fine-tuning acoustic models for AD embedding extraction.

3.2 Model Performance

In total, 2,047 different feature set combinations were used to train the 10 different classifiers and regressors listed in Table 8. For both tasks two different Random Forest models were used, one using 16 trees and the other 100 trees. Model performance evaluation was performed on the development set, as well as using 5-fold stratified cross-validation across 9 different seeds. While this resulted in a large volume of results (four different result sets, each with size 22,517), the final evaluation was performed on a held-out test set, and as per the PROCESS challenge rules, only 6 submissions in total could be made (3 for the classification task and 3 for the regression task).

Table 8. Classification and regression models for diagnosis and MMSE prediction.

Diagnosis Models	**MMSE Prediction Models**
Random Forest (RF)	Random Forest Regressor (RFR)
Logistic Regression (LR)	Lasso Regression (Lasso)
Decision Tree (DT)	Decision Tree Regressor (DTR)
XGBoost (XGB)	XGBoost Regressor (XGBR)
Support Vector Machine (SVM)	Support Vector Regressor (SVR)
Naive Bayes (NB)	Linear Regression (LIR)
Gradient Boosting (GB)	Gradient Boosting Regressor (GBR)
K-Nearest Neighbors (KNN)	K-Nearest Neighbors Regressor (KNNR)
Multilayer Perceptron Neural Network (MLP)	Multilayer Perceptron Regressor (MLPR)
Voting Classifier (VC)	Ridge Regression (RR)

3.2.1 Development Set Results

The baseline results for the development set provided by the PROCESS challenge organizers[2] are presented in Table 9 and Table 10. Table 9 shows classification performance for the Cookie Theft, semantic/phonemic fluency (VF), and their combination (CTD+VF) using accuracy, macro-precision, macro-recall, and macro-F1 score.

Table 9. Baseline results for the diagnosis task (dev).

Model	Prompt	Acc.	Prec.	Rec.	F1
SVM (eGeMAPS)	CTD	0.525	0.399	0.390	**0.393**
	VF	0.375	0.333	0.337	0.324
	CT+VF	0.525	0.368	0.397	0.379
RF (eGeMAPS)	CTD	0.450	0.296	0.330	0.312
	VF	0.450	0.292	0.324	0.307
	CTD+VF	0.425	0.271	0.302	0.285
RoBERTa-Classifier	CTD	0.550	0.381	0.368	0.326
	VF	0.525	0.343	0.346	0.292
	CTD+VF	0.500	0.302	0.337	0.301

The highest macro-F1 score of 0.393 was achieved using eGeMAPS features extracted from the Cookie Theft Picture description recording with a SVM classifier. Table 10 presents the MMSE score prediction regression results for the same tasks.

[2] `https://processchallenge.github.io/dataset/`

Table 10. Baseline results for the MMSE prediction task (dev).

Model	CTD	VF	CTD+VF
SVR (eGeMAPS)	3.06	4.19	7.81
RFR (eGeMAPS)	2.82	3.22	3.11
RoBERTa-Regression	2.75	2.75	**2.74**

The lowest RMSE, 2.74, was achived using the RoBERTa-Regression model using transcripts from the Cookie Theft Picture description and verbal fluency recordings.

In the diagnosis task, 52% or 11,750 model–feature set combinations beat the baselines. In the MMSE prediction task, this percentage was 68%, and the number of model–feature set combinations that beat the organizers' baseline was 15,351. The best feature combination for each algorithm is presented in Table 11.

Table 11. Best model–feature set combination pairs for diagnosis (dev).

Algorithm	Combination	F1 Score
GB	mean_gender, real_age, xvec	**0.670**
MLP	lengths, mean_age, pft, ctd, ctdp, xvec	0.659
RF 16	mean_age, mean_emo, pft, ctd, ege, xvec	0.590
LR	mean_emo, sft, xvec	0.560
DT	mean_gender, mean_emo, ctd, ege, ctdp	0.558
RF 100	lengths, mean_gender, ctd	0.556
NB	lengths, mean_gender, mean_emo, sft, ctdp	0.550
XGB	mean_age, sft, ctd, real_age, ctdp, xvec	0.549
KNN	lengths, mean_age, mean_emo, pft	0.539
VC	mean_emo, ege, real_age, ctdp, xvec	0.520
SVM	lengths, mean_gender, pft, sft, ctd, real_age	0.505
(Baseline) SVM	eGeMAPS (CTD)	0.393

The Gradient Boosting classifier and Multi-Layer Perceptron achieve the highest macro-F1 scores—0.670 and 0.659, respectively—significantly outperforming the baseline results. Interestingly, the Support Vector Machine, which is frequently reported in the literature as a strong performer, ranks lowest among the evaluated algorithms here. Table 12 summarizes the top-performing model–feature set combinations for the MMSE prediction task.

Table 12. Best model–feature set combination pairs for MMSE (dev).

Algorithm	Combination	RMSE
DTR	pft, sft, ege	**1.66**
XGBR	sft, ege, real_age, ctdp	1.80
GBR	lengths, mean_gender, mean_emo, real_age, ctdp	1.82
RFR 16	lengths, mean_gender, mean_age, mean_emo, ctd, ege, ctdp, xvec	1.87
RR	lengths, sft, ege, real_age	1.96
Lasso	mean_gender, mean_emo, pft, sft, real_age, xvec	2.00
RFR 100	mean_gender, mean_age, sft, real_age, ctdp, xvec	2.01
SVR	mean_gender, mean_emo, sft, real_age, ctdp	2.02
KNN	pft, ctd, real_age, ctdp	2.03
LIR	mean_emo, real_age	2.28
MLP	ctdp	2.44
(Baseline) RoBERTa	CTD + PFT + SFT transcripts	2.74

The Decision Tree Regressor achieves the lowest RMSE of 1.66, surpassing the baseline by a noticeable margin. The XGBoost, Gradient Boosting, and Random Forest regressors also show strong performances. For both the diagnosis and MMSE prediction tasks, at least one feature set combination outperformed the baseline results for each algorithm.

3.2.2 Cross-validation Results

The PROCESS challenge organizers did not provide results for cross-validation; instead, the best performance on the development set is used for comparison, which may favor their results, as cross-validation scores typically tend to be lower. The top feature combination for each algorithm in the diagnosis task is presented in Table 13.

Table 13. Best model–feature set combination pairs for diagnosis (CV).

Algorithm	Combination	Mean F1 Score
MLP	mean_emo, ctdp, xvec	**0.511 ± 0.090**
SVM	mean_emo, pft, ege, xvec	0.481 ± 0.087
XGB	mean_gender, mean_age, mean_emo, pft, sft, real_age, xvec	0.470 ± 0.076
NB	mean_age, mean_emo, ctd, ege, real_age	0.462 ± 0.089
LR	mean_gender, ege, real_age, xvec	0.460 ± 0.081
VC	mean_gender, real_age, xvec	0.450 ± 0.071
RF 16	mean_gender, mean_age, ege, xvec	0.448 ± 0.079
RF 100	mean_gender, mean_age, sft, ctd, ege	0.447 ± 0.090
KNN	lengths, mean_gender, mean_emo, pft, ege, real_age	0.445 ± 0.070
GB	mean_gender, pft, sft, ctd, ege, real_age, ctdp, xvec	0.445 ± 0.073
DT	lengths, mean_gender, pft, sft, ctd, ctdp, xvec	0.421 ± 0.089
(Baseline) SVM	eGeMAPS (CTD)	0.393

The macro-F1 scores are on average noticeably lower compared to development set performance. The top 22 models are all MLP and Support Vector Machine has now also emerged as a top-performer. The best feature combination for each algorithm for the MMSE prediction task is presented in Table 14.

Table 14. Best model–feature set combination pairs for MMSE (CV).

Algorithm	Combination	Mean RMSE
Lasso Regression	pft, sft, real_age	**2.076 ± 0.327**
SVR	pft, ctd, real_age, ctdp	2.088 ± 0.366
KNN	mean_age, pft, sft, ctd, real_age, ctdp	2.138 ± 0.455
RFR 100	lengths, mean_gender, pft, real_age	2.148 ± 0.434
RFR 16	lengths, pft, real_age	2.165 ± 0.464
RR	pft, real_age, ctdp	2.169 ± 0.334
GBR	lengths, mean_gender, pft, real_age	2.189 ± 0.478
LIR	sft, real_age	2.225 ± 0.278
MLP	mean_gender	2.324 ± 0.562
XGBR	lengths, mean_gender, real_age, ctdp	2.447 ± 0.470
DTR	mean_gender	2.494 ± 0.554
(Baseline) RoBERTa	CTD + PFT + SFT transcripts	2.74

Lasso and SVR achieve the best mean RMSE scores, 2.076 and 2.088, respectively, while the DTR experiences the most significant drop in ranking compared to its performance on the development set. Figures 17 and 18 present the percentages of model–feature set combinations per algorithm which beat the baselines.

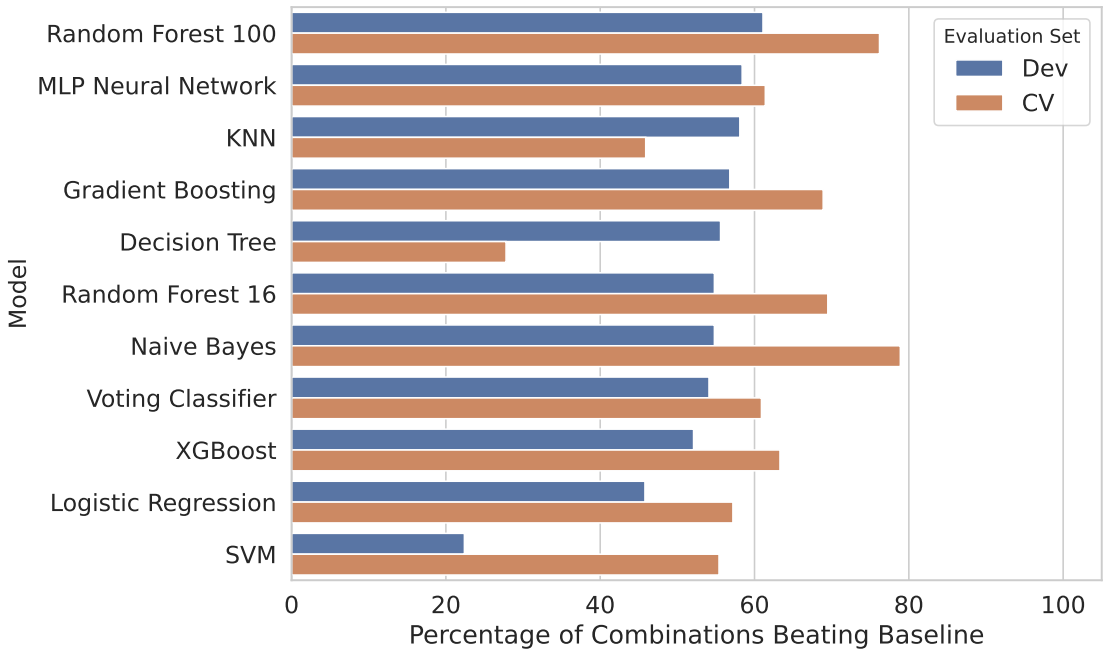

Figure 17. Percentage of model–feature set combinations beating baselines (diagnosis).

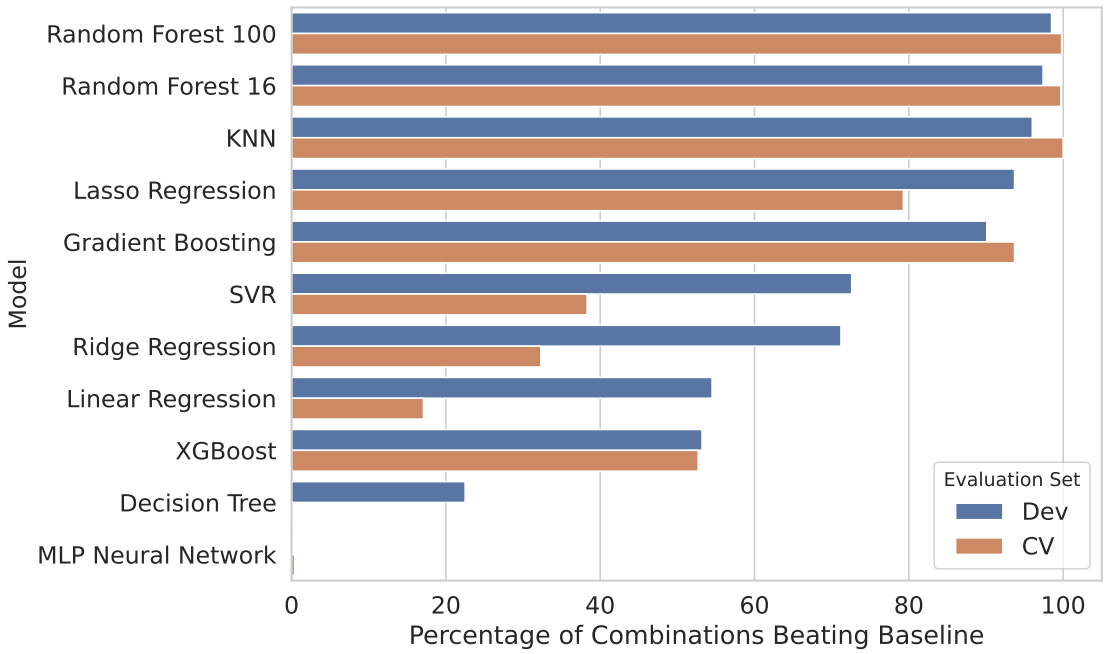

Figure 18. Percentage of model–feature set combinations beating baselines (MMSE).

Figure 19 provides an overview of the distribution of algorithms in the top-N best models by performance for the diagnosis task. The top 10 models are dominated by MLP. The proportion of MLP decreases, reaching near parity with SVM at rank 100, with a few instances of XGBoost. Starting from rank 1000, Naive Bayes is increasingly represented, followed by other models eventually.

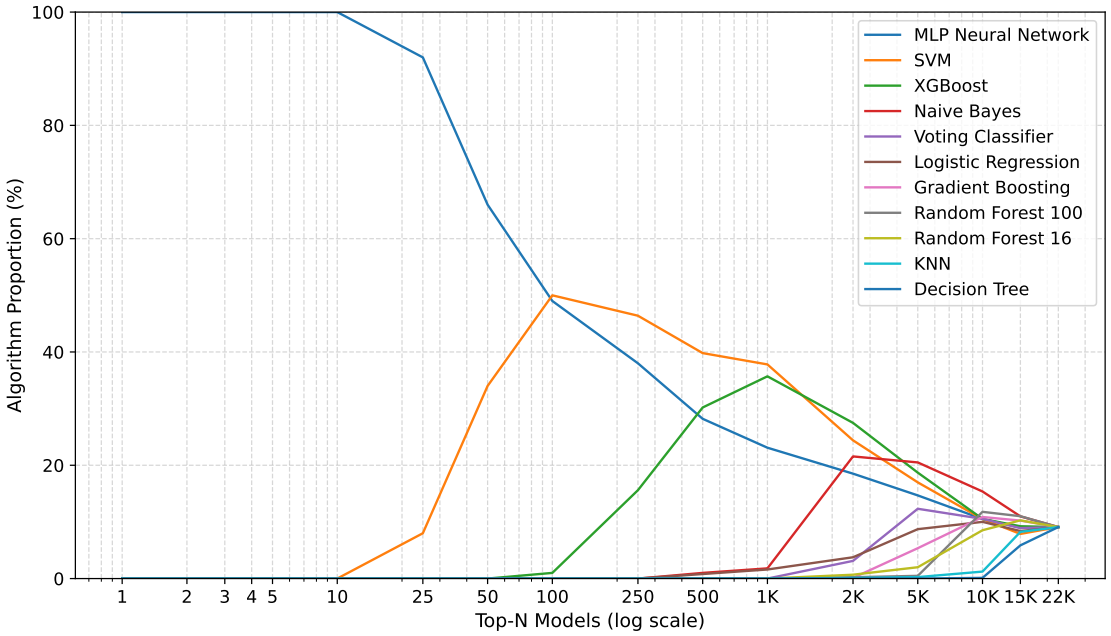

Figure 19. Algorithm distribution trends in top-N diagnosis models.

Figure 20 shows the algorithm trends for the MMSE prediction task, with the top 500 being dominated by Lasso and to a lesser extent SVR. As also seen in Figure 18, both variants of Random Forest Regressor very often beat baselines with different feature combinations, but fail to achieve the lowest RMSE scores.

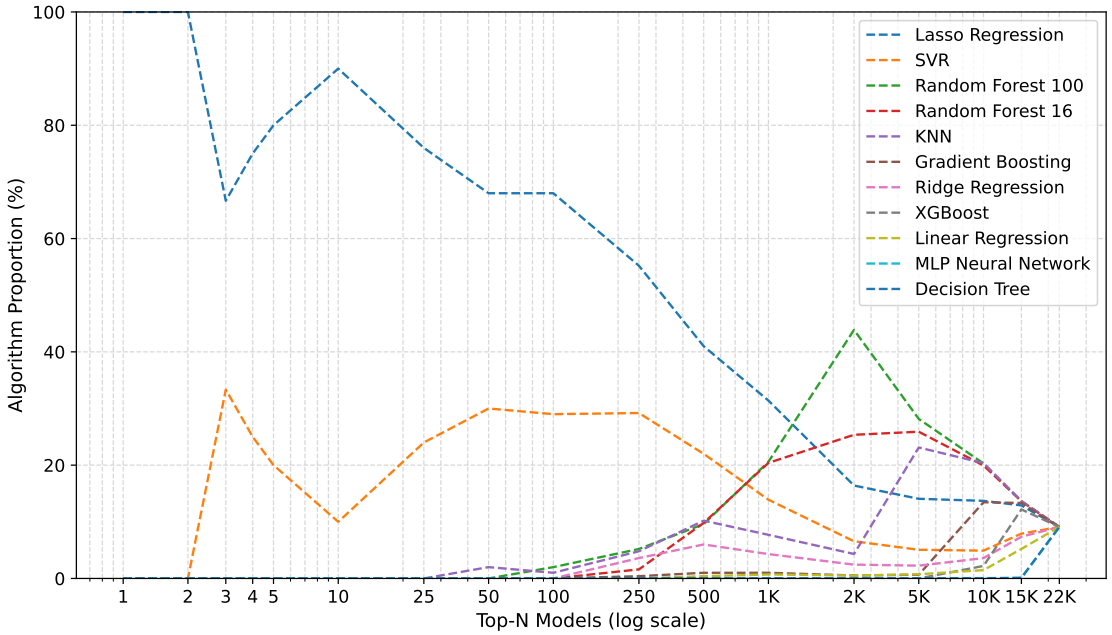

Figure 20. Algorithm distribution trends in top-N MMSE models.

To summarize the cross-validation and development set performance across all model–feature set combinations, Tables 15 and 16 provide an aggregated overview of the most prominent algorithms for the classification and MMSE regression tasks, respectively. The tables report, for each algorithm, the proportion of models that beat the baseline ("Dev Baseline" and "CV Baseline"), the percentage contribution of the algorithm to all baseline-surpassing results ("Dev Contrib." and "CV Contrib."), the highest ranking achieved both when evaluating on the development set as well as in cross-validation setting ("Dev Rank." and "CV Rank.") and the best observed macro-F1/RMSE scores in each setting. These results are discussed in detail in Section 4.1.

Table 15. Summary of classification performance across models.

Algorithm	Dev Baseline (%)	CV Baseline (%)	Dev Contrib. (%)	CV Contrib. (%)	Dev Rank	CV Rank	Dev Best F1	CV Best F1
MLP	58.4	61.4	10.2	9.2	3	1	0.659	0.511
SVM	22.4	55.3	3.9	8.3	1620	22	0.505	0.481
XGB	52.1	63.3	9.1	9.5	165	99	0.549	0.470
NB	54.8	78.9	9.5	11.9	163	304	0.550	0.462
LR	45.8	57.2	8.0	8.6	127	436	0.560	0.460
VC	54.1	60.9	9.4	9.2	929	1189	0.520	0.450
RF 16	54.8	69.5	9.5	10.4	39	1379	0.590	0.448
RF 100	61.1	76.2	10.6	11.4	148	1572	0.556	0.447
KNN	58.1	45.9	10.1	6.9	359	1874	0.539	0.445
GB	56.8	68.8	9.9	10.3	1	1899	0.670	0.445
DT	55.6	27.8	9.7	4.2	139	7954	0.558	0.421

Table 16. Summary of MMSE regression performance across models.

Algorithm	Dev Baseline (%)	CV Baseline (%)	Dev Contrib. (%)	CV Contrib. (%)	Dev Rank	CV Rank	Dev Best RMSE	CV Best RMSE
Lasso	93.6	79.2	12.5	12.9	159	1	1.998	2.076
SVR	72.6	38.3	9.7	6.3	234	3	2.018	2.088
KNNR	96.0	100.0	12.8	16.3	314	43	2.034	2.138
RF 100	98.5	99.8	13.1	16.3	206	62	2.013	2.148
RF 16	97.4	99.7	13.0	16.2	23	104	1.866	2.165
RR	71.2	32.3	9.5	5.3	96	116	1.962	2.169
GBR	90.1	93.6	12.0	15.3	14	166	1.824	2.189
LIR	54.5	17.1	7.3	2.8	4349	291	2.283	2.225
MLP	0.2	0.4	0.0	0.1	9022	3132	2.438	2.324
XGBR	53.2	52.7	7.1	8.6	6	6658	1.801	2.447
DTR	22.5	0.1	3.0	0.0	1	7650	1.658	2.494

3.2.3 Test Set Results

The baseline results for the test set provided by the PROCESS challenge organizers [50] are presented in Table 17 and Table 18. These results were calculated without any hyperparameter optimization. Table 9 shows classification performance for the Cookie Theft, semantic/phonemic fluency (VF), and their combination (CTD+VF) using accuracy, macro-precision, macro-recall, and macro-F1 score.

Table 17. Baseline results for the diagnosis task (test).

Model	Prompt	Acc.	Prec.	Rec.	F1
SVM (eGeMAPS)	CTD	0.575	0.535	0.612	**0.550**
	VF	0.450	0.388	0.391	0.383
	CTD+VF	0.500	0.417	0.423	0.417
RF (eGeMAPS)	CTD	0.600	0.717	0.499	0.533
	VF	0.525	0.331	0.359	0.339
	CTD+VF	0.525	0.661	0.439	0.474
RoBERTa-Classifier	CTD	0.525	0.361	0.397	0.368
	VF	0.550	0.356	0.381	0.356
	CTD+VF	0.525	0.322	0.359	0.329

In the classification task, the best baseline F1-score of 0.550 was achieved using an SVM classifier with eGeMAPS features extracted from the CTD task audio. This represents a substantial improvement over the development set score of 0.393, which may suggest distributional differences between the datasets or reflect the capacity of eGeMAPS features to benefit from a larger amount of data. Alternatively, the improvement could stem from correctly predicting instances of a minority class, which can significantly impact the F1-score in the presence of class imbalance. Table 18 presents the corresponding regression results for the same tasks.

Table 18. Baseline results for the MMSE prediction task (test).

Model	CT	VF	CT+VF
SVR (eGeMAPS)	4.40	7.93	6.68
RFR (eGeMAPS)	3.31	3.31	3.17
RoBERTa-Regression	2.99	2.98	**2.98**

In the MMSE prediction task, the lowest RMSE of 2.98 was achieved by the RoBERTa-Regression model using transcripts from the Cookie Theft picture description and verbal fluency recordings. This performance is slightly worse than the development set result of 2.74. Table 19 and Table 20 present our test set results for the diagnosis and MMSE prediction tasks respectively.

Table 19. Classification model performance (test).

Algorithm	Combination	F1
XGB	lengths, mean_age, pft, ctd, ctdp	**0.609**
MLP	lengths, mean_gender, mean_age, mean_emo, pft, sft, ctd, ege, ctdp, xvec	0.576
RF 16	mean_gender, mean_age, mean_emo, ege, real_age, xvec	0.457
(Baseline) SVM	eGeMAPS (CTD)	0.550

In the classification task, the XGBoost model, using features such as transcript lengths, mean predicted age, phonemic fluency, macro-descriptors, and CTD pause features, achieved the highest macro-F1 score of 0.61. The RF 16 and MLP models, which incorporated a broader set of features including emotional attributes, macro-descriptors, and speaker embeddings, achieved macro-F1 scores of 0.46 and 0.58 on the test set, respectively. It is worth noting that the selection of the MLP model was based on its strong performance in cross-validation, whereas this was not the case for XGB and RF 16. Interestingly, models that performed best during cross-validation tended to produce an unusually low number of predictions for the minority class when evaluated on the test set. To address this issue, the best-performing regressors were used to predict MMSE scores, which were then converted into pseudo-labels. Two classification models were selected based on their macro-F1 scores in predicting these labels.

Table 20. Regression model performance (test).

Algorithm	Combination	RMSE
Lasso	pft, sft, real_age	**2.54**
KNNR	lengths, mean_gender, mean_age, sft	2.88
SVR	pft, sft, mean_age	2.97
(Baseline) RoBERTa	CTD + PFT + SFT transcripts	2.98

For the regression task, Lasso model, using phonemic fluency, semantic fluency, and real age, achieved the best performance with an RMSE of 2.54 on the test set. SVR and KNNR, using different combinations of text-based and demographic features, also performed well with RMSE values of 2.97 and 2.88 on the test set. Notably, acoustic features, such as speaker embeddings and emotional attributes, performed better for classification, while text-based features, such as fluency measures, were more effective for regression.

Our team achieved 3rd place overall in the PROCESS challenge, with a 2nd-place finish in the regression task and a 5th-place finish in the classification task, out of 36 participating teams worldwide, demonstrating competitive performance in this domain.

3.3 Feature Analysis

To better understand which features were most common among the feature set combinations that outperformed the challenge baselines, a more fine-grained analysis was conducted. Figures 21 and 22 present the various feature sets, ordered by their frequency of occurrence in diagnosis and MMSE prediction models that exceeded the development set baselines.

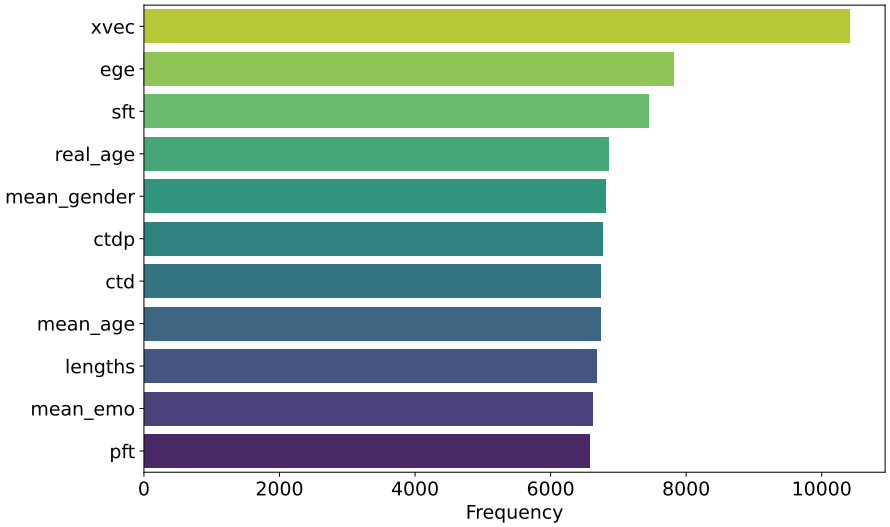

Figure 21. Occurrence frequency of feature sets beating baselines (diagnosis).

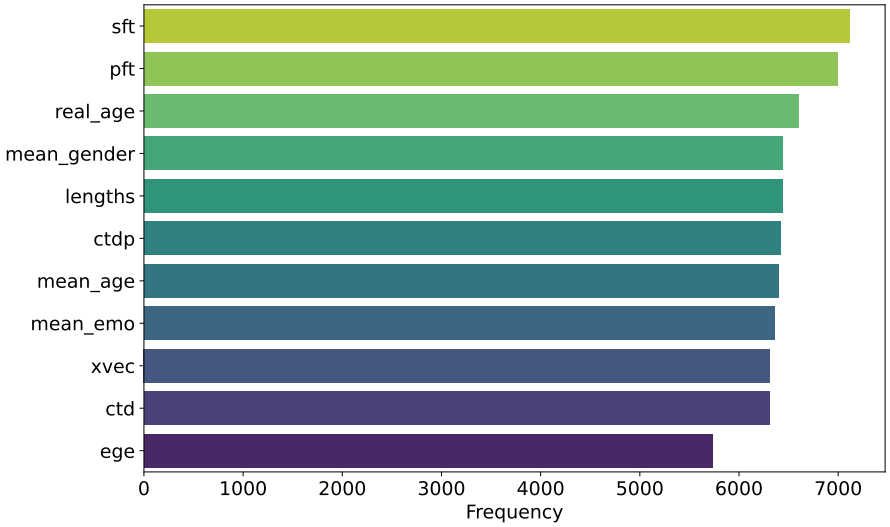

Figure 22. Occurrence frequency of feature sets beating baselines (MMSE).

In general, for the diagnosis models, the most frequently used feature set is the `xvec` AD embeddings, followed by `ege` — both acoustic features. For the MMSE prediction models, fluency features and demographic features are the most frequently used features.

While this information is useful, it does not directly explain whether the frequently used feature sets were also the ones which achieved the highest performance. Figures 23 and 24 present a better overview of the frequency of occurrence of feature sets in top-N models.

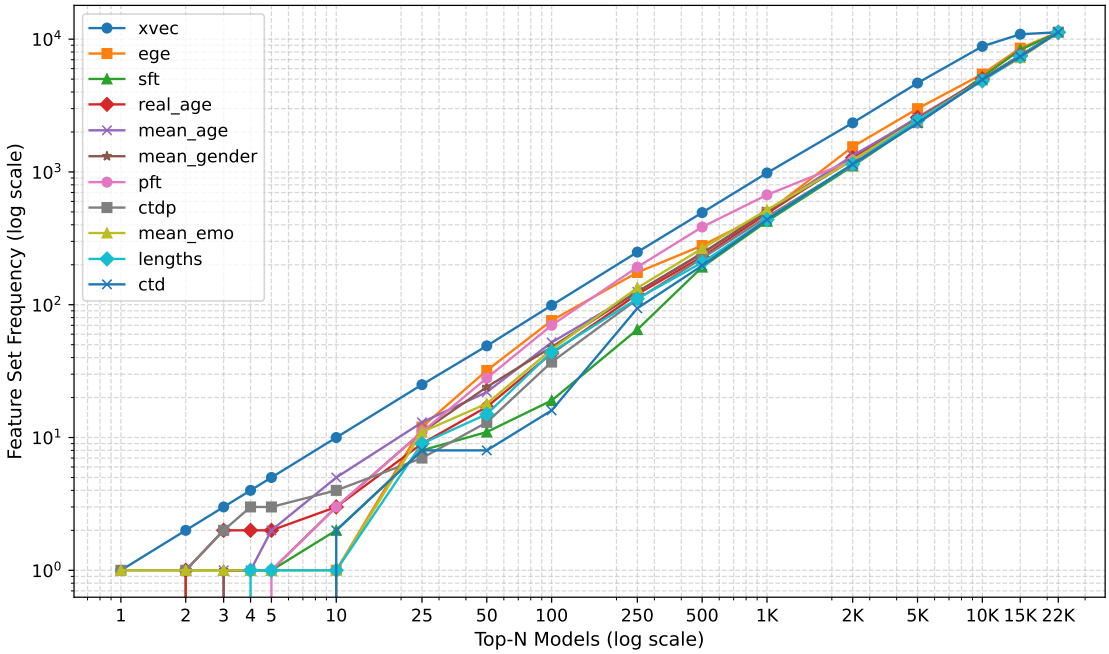

Figure 23. Top feature set frequencies in diagnosis models.

For the diagnosis task, `xvec` dominates the top feature sets throughout.

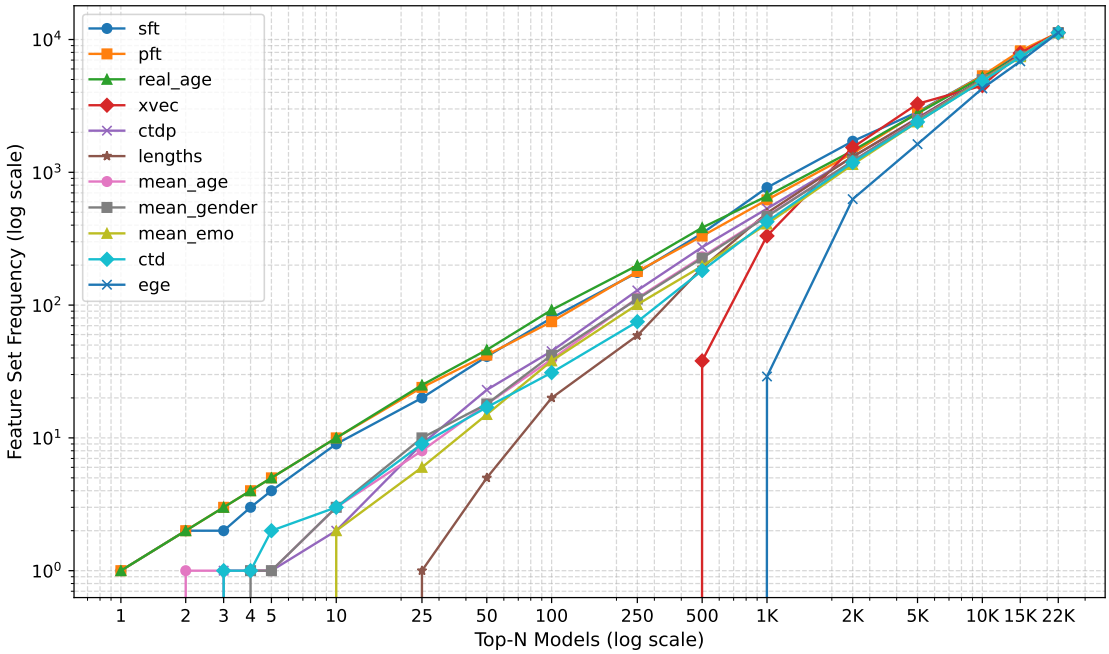

Figure 24. Top feature set frequencies in MMSE models.

Interestingly, for the MMSE prediction task, acoustic feature sets drop out relatively early, with `ege` not appearing even once in the top 690 results, and `xvec` not reaching the top 320. Further analysis was conducted to understand the trends of feature sets occurrences. For this, the features were categorized in two ways. First, based on whether the feature sets were acoustic, text-based or demographic. Secondly, based on the extraction type of the features, e.g., *latent* features represent embedding vectors derived from deep learning models, such a `xvec`, *semantic* features represent features which are also derived from deep learning models but that have a meaning attached, e.g., `mean_emo`, `mean_age`, `mean_gender`, and `ctd`, and *concrete* features, which are extracted programmatically and are fully interpretable.

In addition to presenting the raw proportions of each feature type, the weighted proportions are also presented since the number of feature sets belonging to the different categories are not equal. Figure 25 and Figure 26 present the raw feature category proportions for both the diagnosis and MMSE prediction tasks.

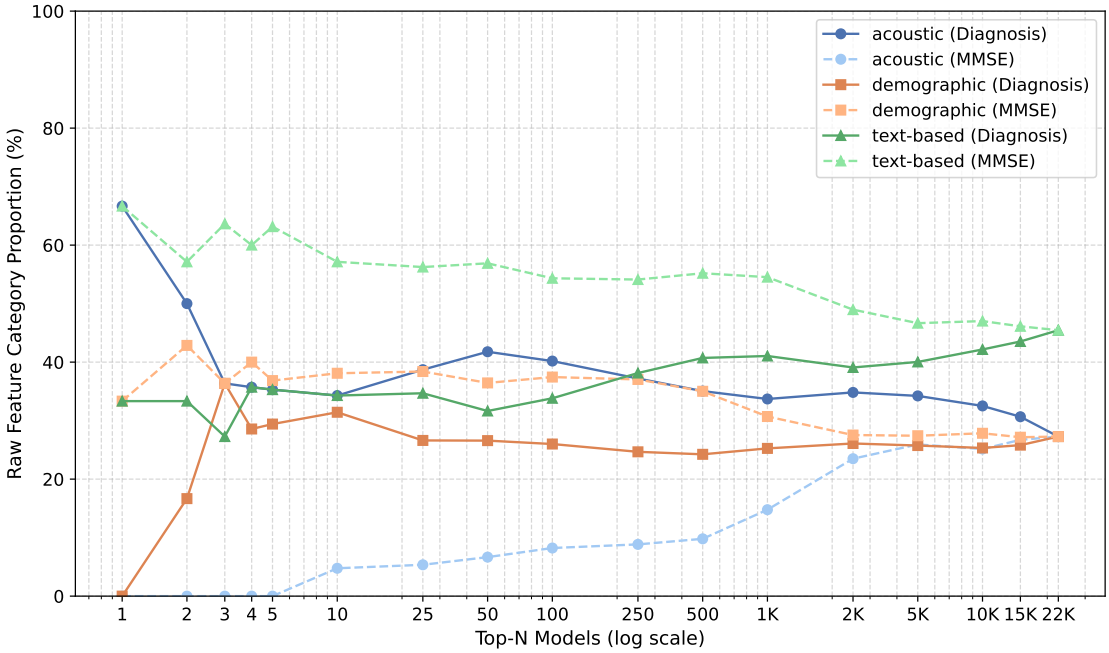

Figure 25. Raw proportions of feature types across Top-N models (data type).

Based on their frequency of occurrence, models for the MMSE prediction task rely more heavily on text-based and demographic features, whereas classifiers for the diagnosis task tend to place greater emphasis on acoustic features.

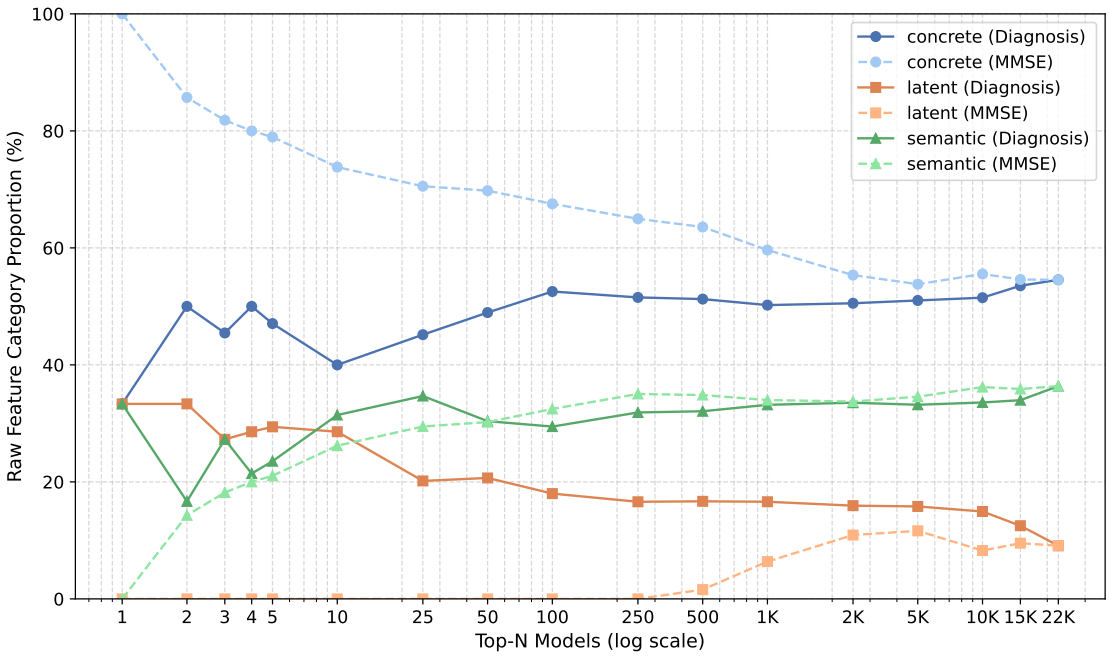

Figure 26. Raw proportions of feature types across top-N models (extraction type).

When grouping features by their extraction type, we observe that regression models increasingly favor concrete features among the top-performing configurations, whereas classifiers show a growing reliance on latent features.

Figure 27 and Figure 28 present the weighted feature category proportions for both tasks.

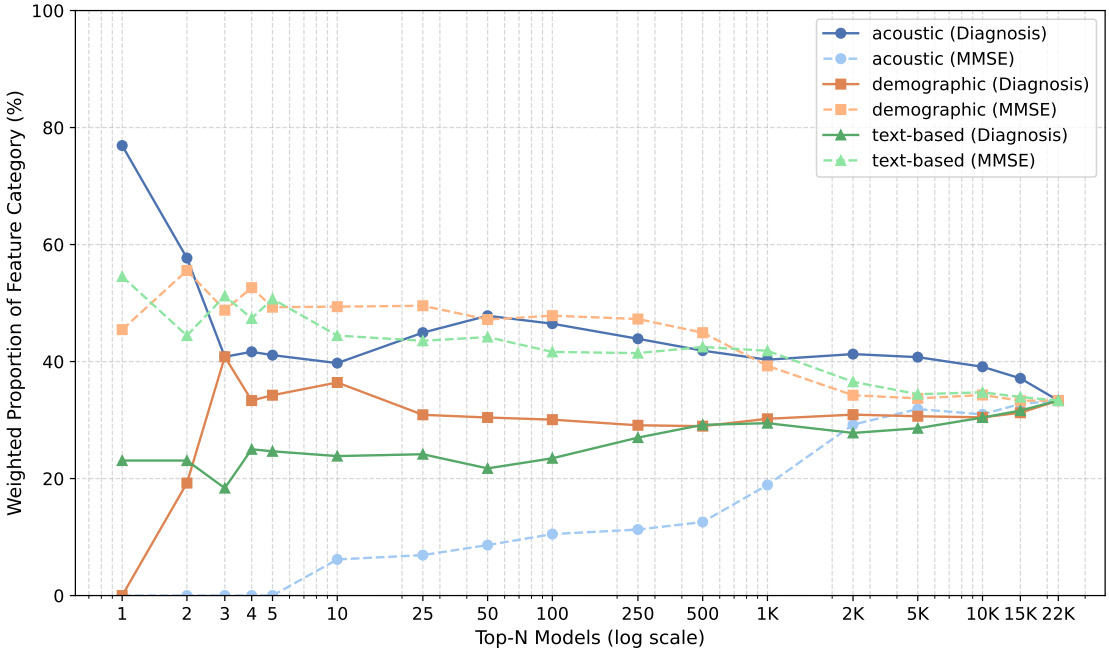

Figure 27. Weighted proportions of feature types across top-N models (data type).

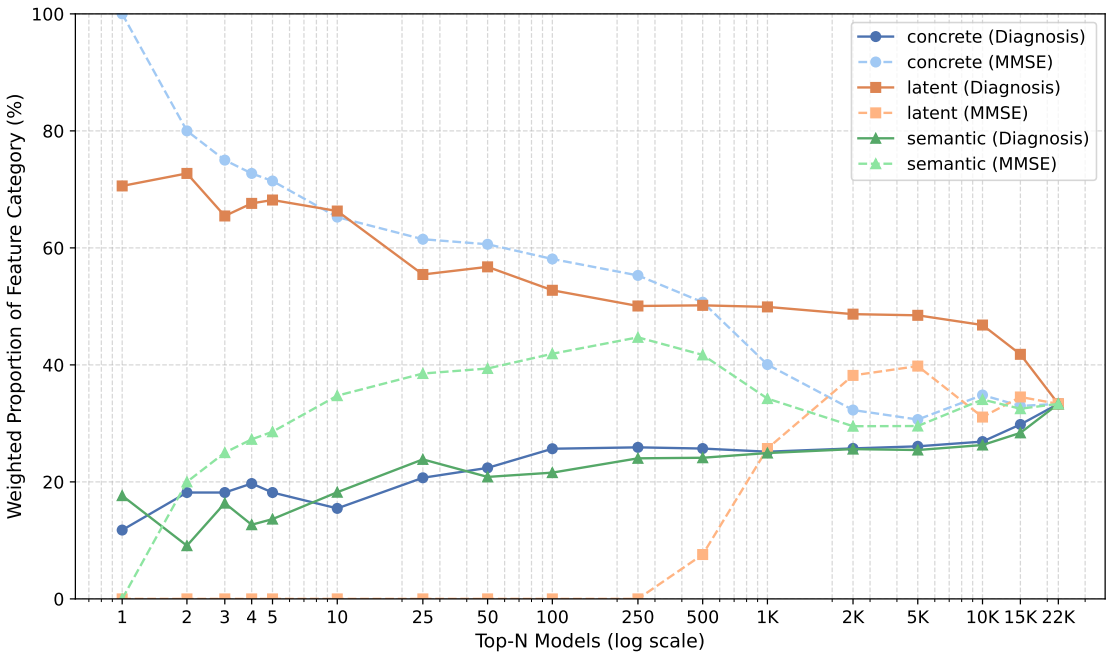

Figure 28. Weighted proportions of feature types across top-N models (extraction type).

It can be noted, that for the diagnosis task the acoustic features are consistently the most frequently occurring, followed by demographic and text-based features last. In MMSE prediction this trend is reversed, with demographic and text-based features performing better, and acoustic features less so. For the diagnosis task, latent features clearly dominate among the top-performing features. In contrast, for MMSE prediction, concrete features perform best, followed by semantic features, while latent features appear to be less useful.

Figures 29 and 30 present the raw proportions of feature types for both methods of categorization for each task separately.

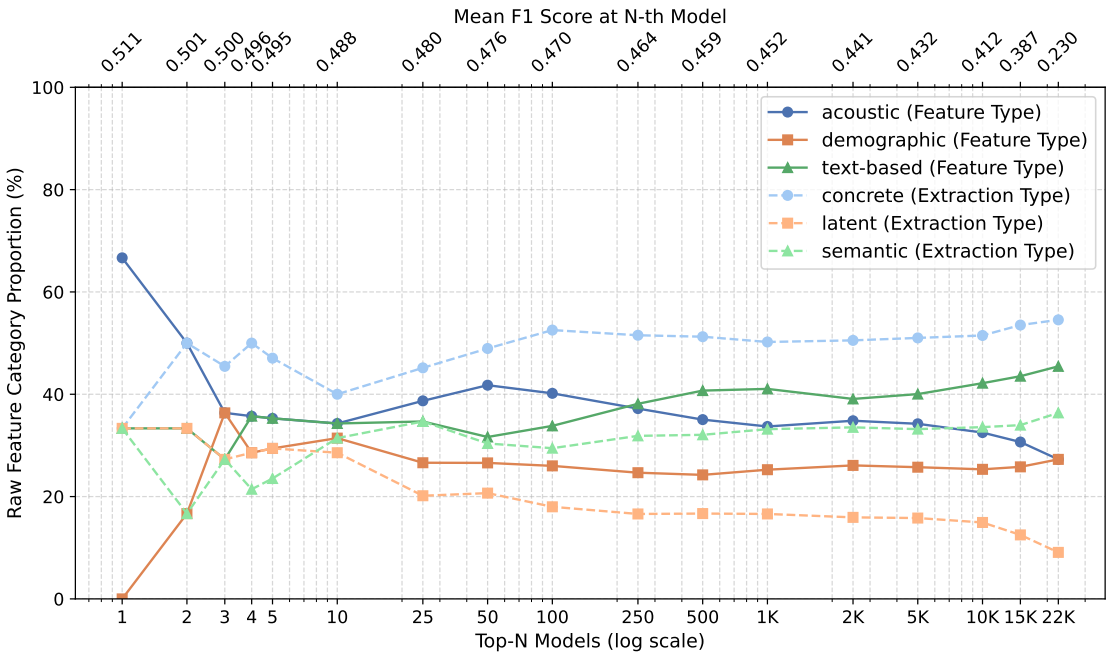

Figure 29. Diagnosis: Raw proportions of feature types across top-N models.

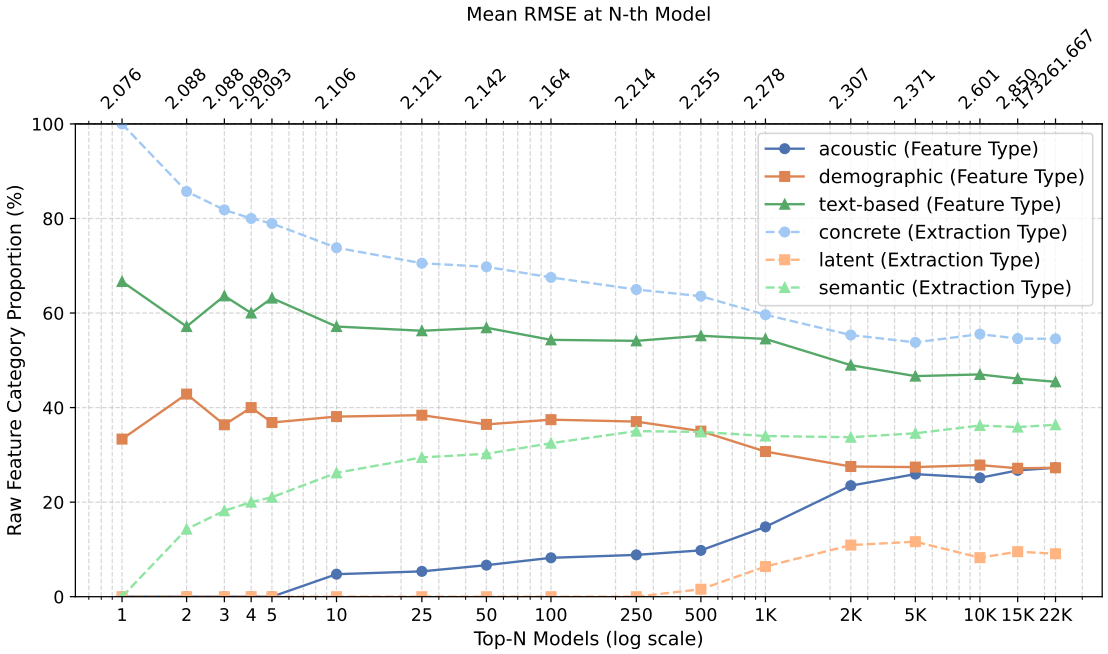

Figure 30. MMSE: Raw proportions of feature types across top-N models.

Finally, Figures 31 and 32 present the weighted proportions of feature types for both methods of categorization for each task separately, providing clearest insights.

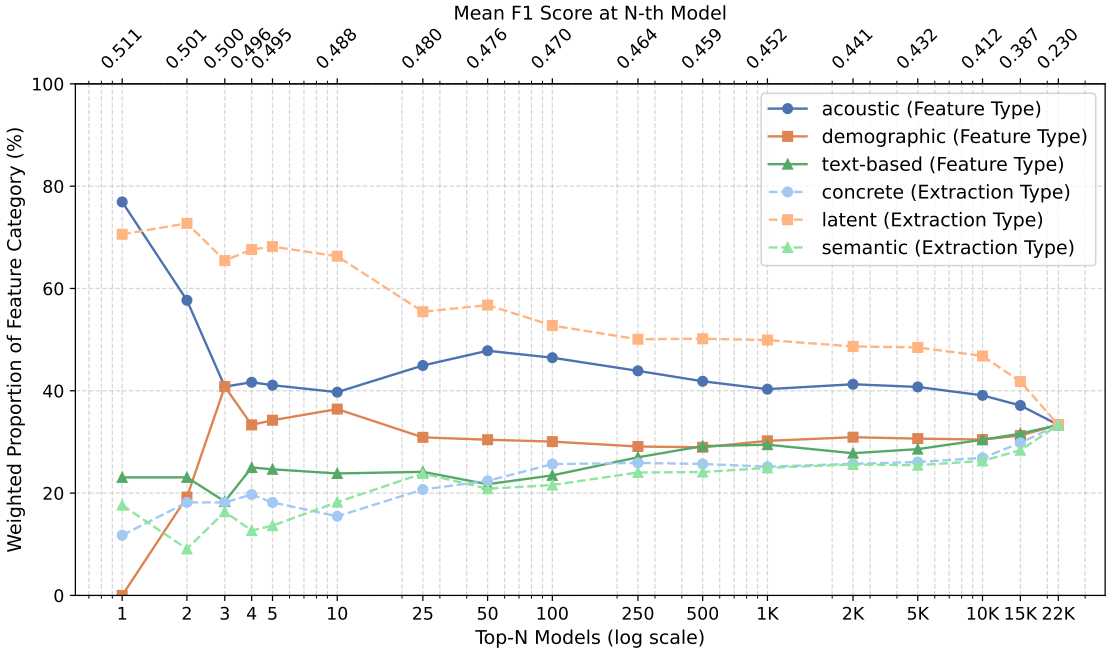

Figure 31. Diagnosis: Weighted proportions of feature types across top-N models.

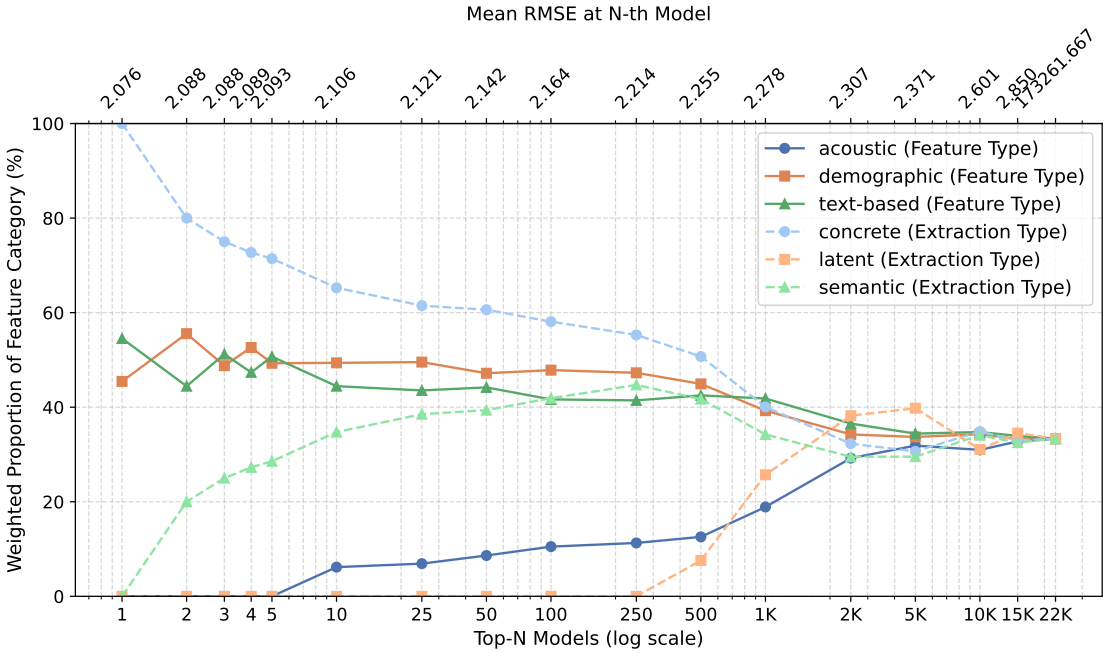

Figure 32. MMSE: Weighted proportions of feature types across top-N models.

Latent acoustic features dominate the diagnostic task, while text based and demographic features are more important for MMSE prediction, and these features tend to be concrete.

3.4 Interpretability Insights

This subsection provides interpretability insights through the analysis of decision tree splits, global SHAP feature importances, Lasso coefficients, and Pareto fronts that highlight the trade-off between model performance and the Joint Interpretability Index.

3.4.1 MMSE Predictions via Decision Tree Logic

To enhance interpretability and clinical relevance, the structure of a regression decision tree trained to predict MMSE scores from acoustic, demographic, and linguistic features derived from speech, was analyzed. The tree achieved a RMSE of 1.658 on the development set. The DTR splits are provided in Appendix 2. Below, a high-level interpretation of the model logic is provided, highlighting potential relationships between the features and predicted cognitive status.

Primary Split: Semantic Fluency as a Cognitive Indicator

The first and most influential decision node is based on `animals_sft`, the number of animals named in 60 seconds. A threshold of 12.5 words serves as the primary split:

- **Low fluency ($\leq$ 12.5)**: Suggests cognitive impairment (Mild Cognitive Impairment or Dementia).
- **Higher fluency ($>$ 12.5)**: More likely to indicate preserved cognitive function, often corresponding to Healthy Controls (HC) or high-functioning MCI.

This split aligns with clinical observations that semantic fluency may be among the earliest and most sensitive indicators of cognitive decline, particularly in Alzheimer's-type dementia.

Low Fluency Branch: Characterizing Impairment Severity

Among low-fluency participants, the model refines its prediction using prosodic and spectral features:

- `spectralFlux_sma3_mean` captures the variability in the spectral content of speech. Lower values (i.e., more monotonous or flat speech) are common in cognitive decline.
- When spectral flux is low ($\leq$ 0.13), the model further splits on:

  - `mfcc3_sma3_mean`:
    * ≤ 6.65: Predicts MMSE = 22, typically labeled as MCI.
    * > 6.65: Combines with shimmer-based thresholds to predict MMSE = 19–20, usually more consistent with Dementia.
  - `p_words_pft` and `F2bandwidth_sma3nz_mean` help identify slightly less impaired participants (MMSE = 25–26).
- When spectral flux is higher (> 0.13), the model incorporates formant frequencies and temporal dynamics to distinguish MCI (MMSE = 24–26).

High Fluency Branch: Differentiating Healthy Controls

For participants who named more than 12.5 animals, the model uses subtle acoustic markers to distinguish between Healthy Controls and MCI:

- Features such as `mfcc2_sma3_mean`, `count_pauses_sft`, formant amplitudes, and `shimmerLocaldB_sma3nz_std` provide further stratification.
- MMSE predictions in this branch range from 27 to 30:
  - **MMSE = 30**: Assigned to fluent, prosodically rich speakers with low MFCC2 and stable formant dynamics — typically Healthy Controls.
  - **MMSE = 27–28**: Reflects mild prosodic or articulatory irregularities, consistent with high-functioning MCI.
- Pause features (`count_pauses_sft`) also offer discriminative value:
  - Lower pause counts correlate with preserved executive control.
  - Higher pause counts or irregular patterns may indicate cognitive slowing or disorganization.

This decision tree regressor highlights that acoustic features related to voice stability, articulation, and prosody can serve as effective proxies for cognitive function. The hierarchical structure of the model captures both broad and nuanced clinical distinctions. Semantic fluency acts as an early and sensitive screening tool. Voice-based markers help distinguish between MCI and Dementia, as well as between MCI and healthy aging.

3.4.2 SHAP Feature Importances

Below the SHAP feature importances are given for the best model in cross-validation setting scoring 0.51 macro-F1 (Figure 33 and 34). That is, MLP Neural Network with `mean_emo`, `ctdp`, `xvec`. The model relies mainly on latent acoustic features, but also uses some text-based features such as CTD pause features and other semantic acoustic

features like the audio based emotional attributes. For comparison, MLP Neural Network, using only `xvec` achieved 0.41 macro-F1 in cross-validation.

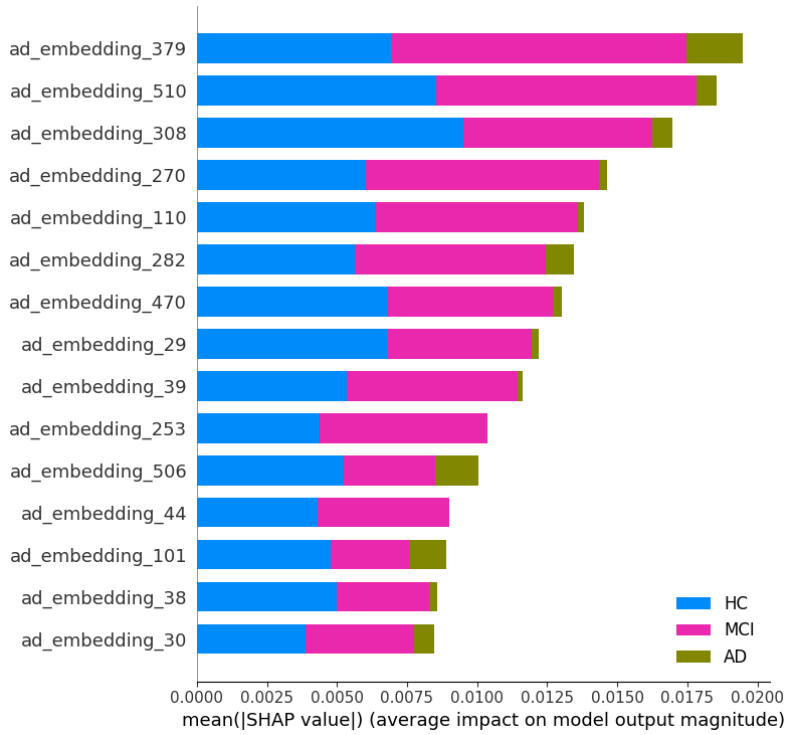

Figure 33. SHAP summary plot for best diagnosis model (CV).

Figure 34 presents the most impactful features per class. In this visualization, blue bars indicate that a given feature contributes positively toward membership in the class, while red bars indicate that higher values of the feature are associated with not belonging to the class. Several features appear to clearly discriminate between HC and MCI (`ad_29`, `ad_271`, `ad_371`), AD and MCI (`ad_506`), and AD and HC (`ad_439`, `ad_480`); interpreting these features in terms of their clinical relevance remains challenging.

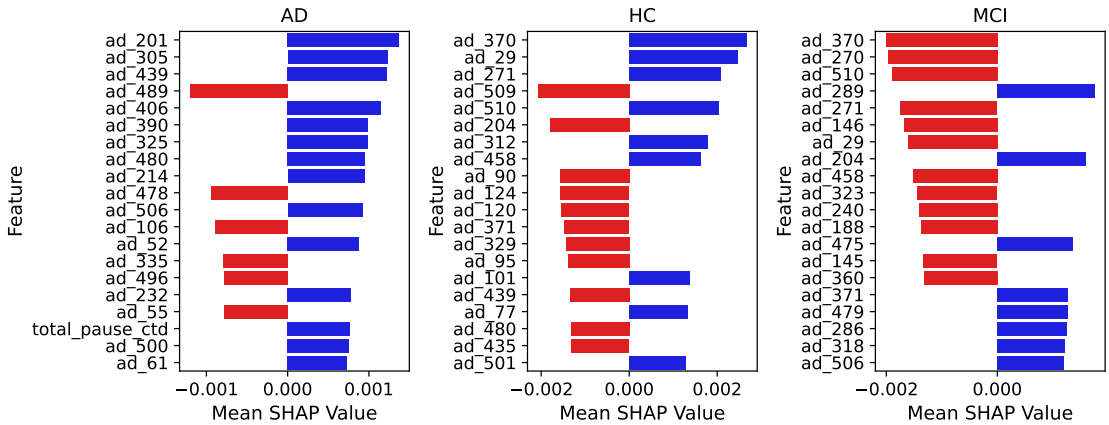

Figure 34. SHAP class-wise feature importances for best diagnosis model (CV).

Figure 35 presents the SHAP summary plot for the top 15 features used by the best-performing test set classifier, XGBoost, trained on the `lengths`, `mean_age`, `pft`, `ctd` and `ctdp` feature sets. In this case it is easier to understand which individual features are most impactful.

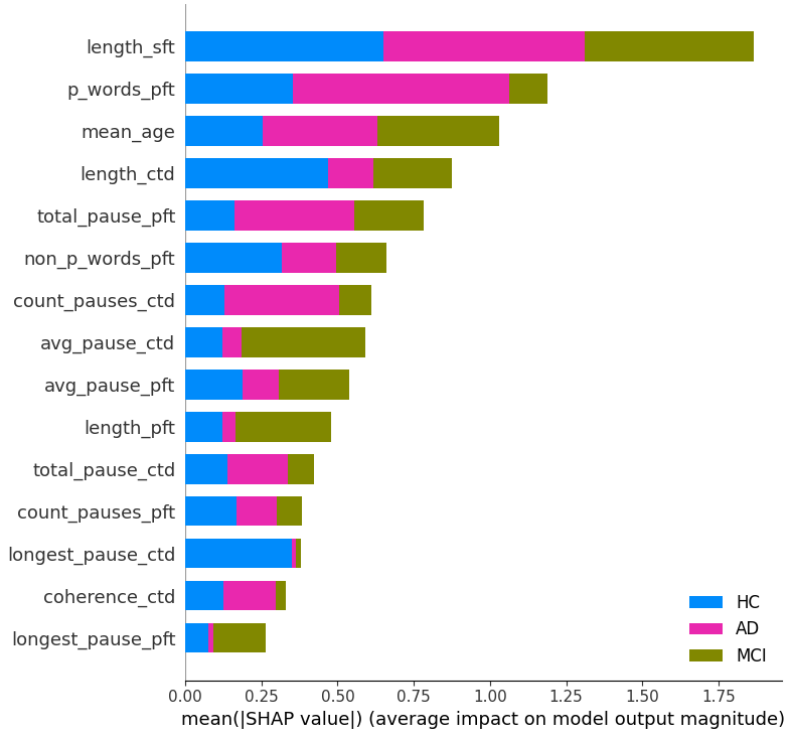

Figure 35. SHAP summary plot for best diagnosis model (test).

Figure 36 gives a class-wise overview of the top features. It can be seen that higher values of P-words and length of SFT decrease the model's confidence that the subject has AD. Surprisingly, here, a higher mean age is also associated with a reduced likelihood of having AD. More intuitively, reduced coherence and increased word finding difficulties hint at possible AD. The biggest indicators of a person belonging to the HC class are the lengths of SFT and CTD transcripts. Interestingly, the more frequent pauses during the CTD task seem to be associated with this class, perhaps signaling reflection, but pause duration in CTD and PFT are negatively associated with this class. For MCI, the main features reducing the likelihood of belonging to this class are lengths of all three tasks as well as mean predicted age.

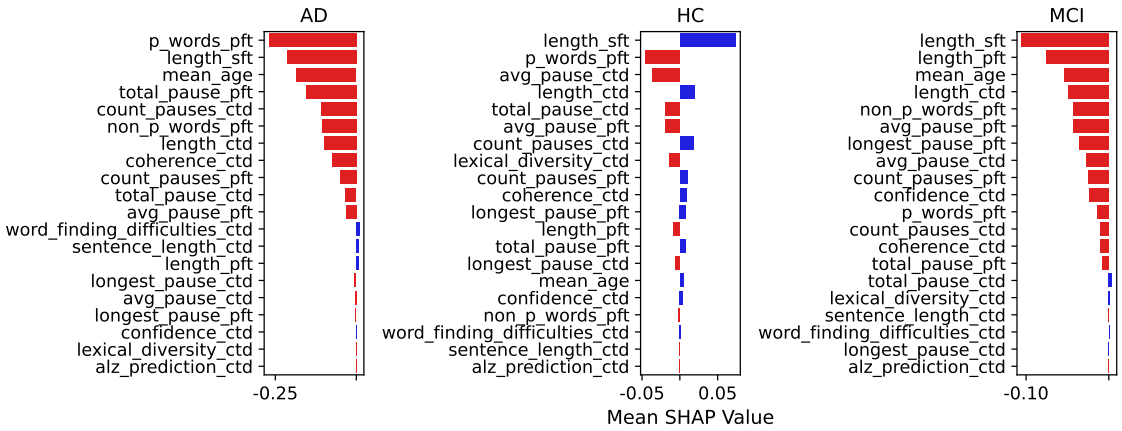

Figure 36. SHAP class-wise feature importances for best diagnosis model (test).

In Figure 37, which presents the SHAP summary plot for the best MMSE prediction model, Lasso, we can see that the number of animals named during the semantic fluency task consistently assign higher MMSE to higher word count. The same is true for words starting with the letter P in the phonemic fluency task. Lower values of real age are associated with higher MMSE scores. Higher pause count in the semantic fluency task and number of words not starting with the letter P are associated with reduction in MMSE score. Interestingly, longest pause in the phonemic fluency task and average pause in the semantic fluency task are associated with slightly higher MMSE scores. Table 21 also provides the coefficients of the Lasso model.

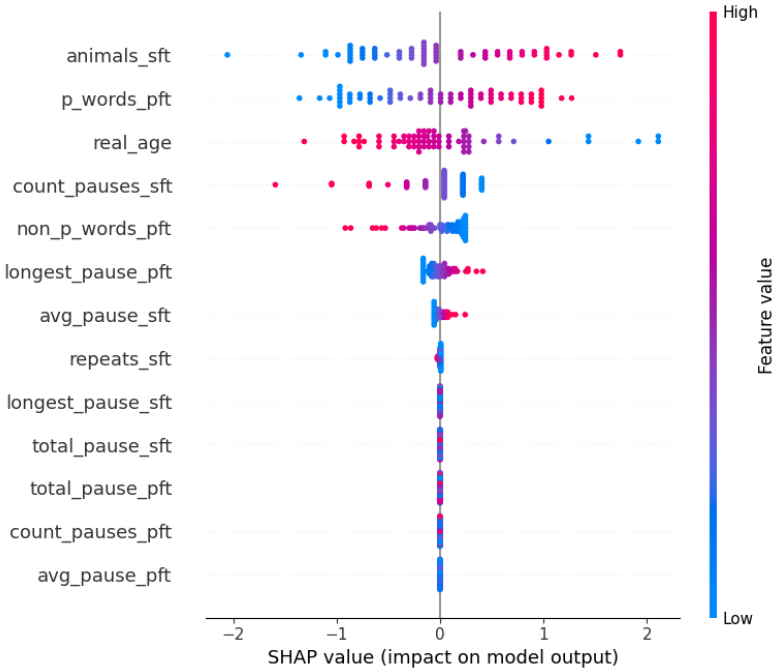

Figure 37. SHAP summary plot for best MMSE model (CV and test).

Table 21. Lasso regression coefficients grouped by sign.

Positive Coefficients		Negative Coefficients	
Feature	Coef.	Feature	Coef.
animals_sft	0.791	count_pauses_pft	-0.000
p_words_pft	0.692	total_pause_sft	-0.000
longest_pause_pft	0.125	repeats_sft	-0.008
avg_pause_sft	0.056	non_p_words_pft	-0.278
avg_pause_pft	0.000	count_pauses_sft	-0.377
longest_pause_sft	0.000	real_age	-0.660
total_pause_pft	0.000		

The Lasso regression model identifies key features that predict cognitive assessment scores (MMSE) based on phonemic and semantic fluency tasks, as well as participants' real age.

- **Positive Predictors:**
  - **Semantic fluency:** The number of animals named (`animals_sft`) is the strongest predictor, with more animals named correlating with higher MMSE scores.
  - **Phonemic fluency:** The number of words starting with the letter P (`p_words_pft`) also shows a positive relationship with cognitive scores.
- **Pause-related Features:**
  - **Pause length:** Both in phonemic and semantic fluency tasks, pause length shows weak positive associations with MMSE, potentially indicating more deliberate speech in cognitively healthy individuals.
  - **Number of pauses:** The number of pauses during the semantic fluency task (`count_pauses_sft`) is negatively correlated with MMSE, suggesting that frequent pauses may indicate cognitive decline.
- **Negative Predictors:**
  - **Non-P words in phonemic fluency:** The number of non-P words (`non_p_words_pft`) and repeated words in semantic fluency (`repeats_sft`) have small negative effects, which could indicate cognitive inefficiency or difficulty in task execution.
- **Real Age:** As expected, age (`real_age`) is negatively correlated with MMSE, reflecting the typical cognitive decline with age.

These results demonstrate the importance of both speech fluency and demographic factors in predicting cognitive health, with semantic fluency being the most predictive feature.

3.4.3 Pareto Fronts

Diagnosis

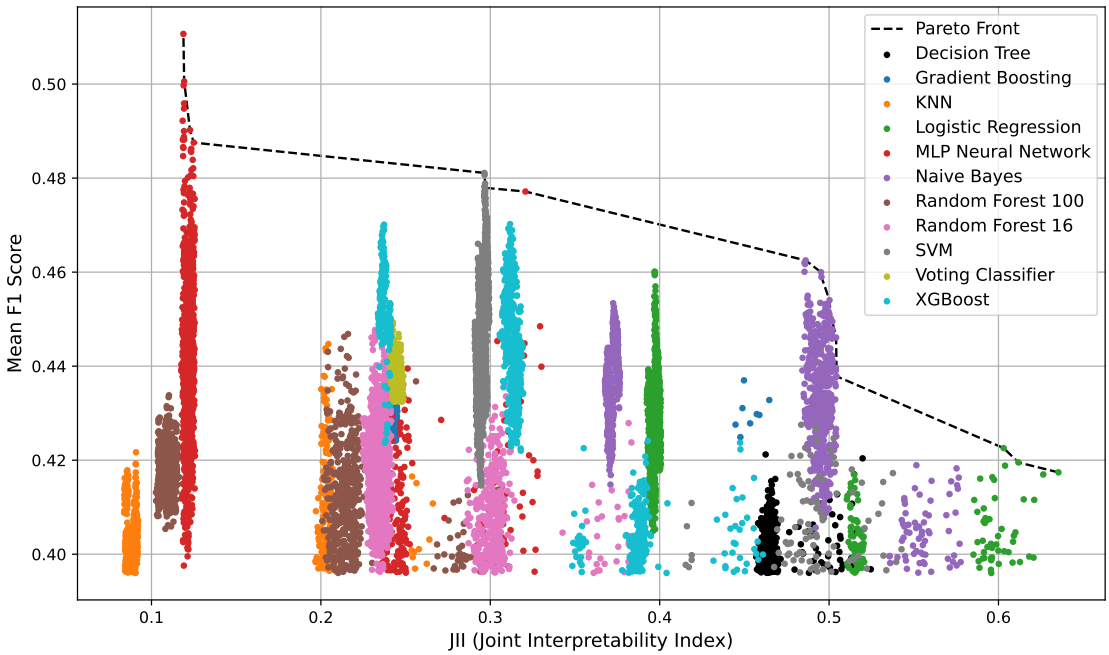

Figure 38. Pareto front: Mean F1 score vs JII.

To investigate the relationship between model interpretability and predictive performance for the diagnosis task (Figure 38), Joint Interpretability Index (JII) scores were computed for each model–feature set combination using equal weights assigned to each of the three main components, with their respective subcomponents also uniformly weighted. Symbolic regression was then applied to approximate the mean macro-F1 score as a function of the JII for models on the Pareto front. The PySR library [51], which performs symbolic regression using evolutionary algorithms, was used to derive an interpretable closed-form expression capturing this relationship.

The input to the model consisted of JII values, while the output was the corresponding cross-validated mean macro-F1 scores. The symbolic regression was configured with a maximum of 1500 iterations, utilizing a set of binary operators $\{+, -, \times, \div, \text{pow}\}$ and unary operators $\{\exp, \log, \sqrt{\cdot}, \text{abs}\}$. Operator complexities were specified to guide the search toward simpler expressions, and constraints were applied to ensure numerical stability (e.g., exponent ranges for `pow`). Model selection was based on a trade-off between simplicity and predictive accuracy, using PySR's `"best"` model selection strategy.

The resulting symbolic expression (3.1) provides an approximation of the relationship

between F1 score and JII, effectively capturing the trade-off behavior observed along the Pareto front.

$$F1 = \left| \left(0.4479027 - \frac{0.2806071}{JII} \right) (JII - 0.16601114) \right| + 0.41417393 \tag{3.1}$$

Figure 39 illustrates the relationship between the JII scores and the corresponding mean macro-F1 scores for a set of models on the Pareto front. The figure shows two curves: one connecting the true F1 scores obtained through cross-validation, and another representing the predicted F1 scores derived from the symbolic regression model given in (3.1). The symbolic model closely approximates the true performance values, with a RMSE of 0.0045. This demonstrates the model's ability to capture the underlying structure of the trade-off between interpretability (as measured by JII) and predictive performance.

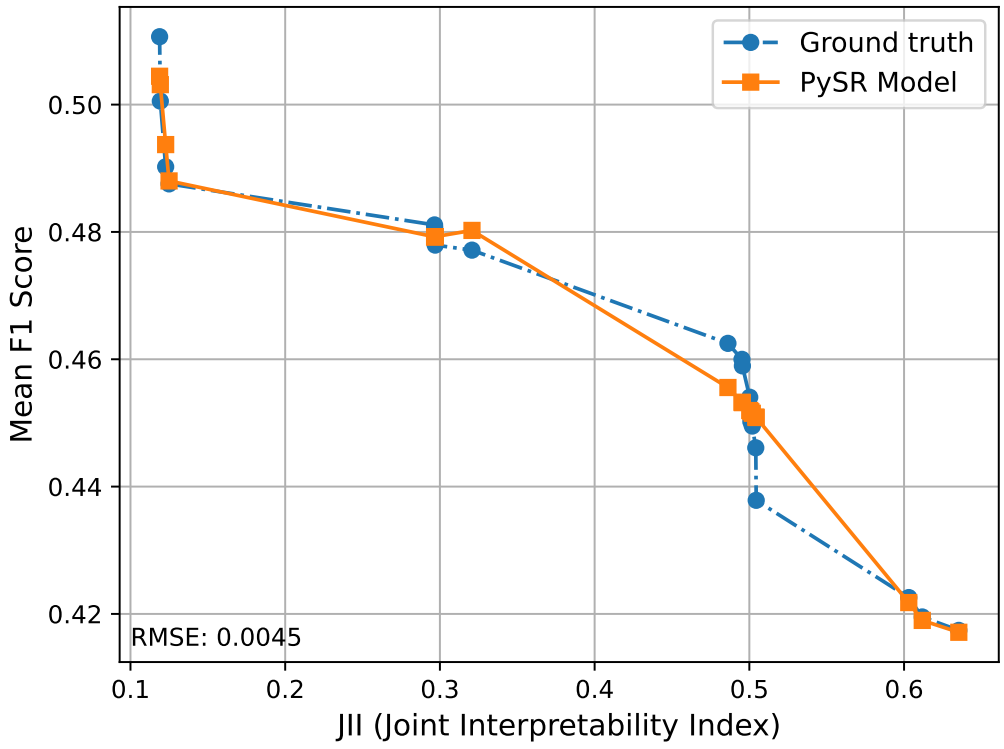

Figure 39. PySR model approximation of Pareto front (diagnosis).

Table 22 presents the mean F1 Score and corresponding JII values for each algorithm combination represented on the Pareto front.

Pareto optimal models, mean macro-F1 vs JII (CV)

Table 22. Pareto optimal models, mean macro-F1 vs. JII (CV).

Algorithm	Combination	Mean F1	JII
MLP	mean_emo, ctdp, xvec	0.511 ± 0.090	0.119
MLP	sft, real_age, xvec	0.501 ± 0.084	0.119
MLP	mean_gender, mean_age, ctd, ege, ctdp, xvec	0.490 ± 0.101	0.123
MLP	lengths, mean_gender, mean_age, mean_emo, pft, sft, ctd, ege, real_age, xvec	0.488 ± 0.114	0.125
SVM	mean_emo, pft, ege, xvec	0.481 ± 0.087	0.297
SVM	mean_age, mean_emo, pft, ege, xvec	0.481 ± 0.085	0.297
SVM	lengths, mean_age, pft, ege, xvec	0.479 ± 0.092	0.297
SVM	lengths, mean_gender, mean_age, pft, ege, xvec	0.478 ± 0.088	0.297
MLP	mean_age, sft	0.477 ± 0.100	0.321
NB	mean_age, mean_emo, ctd, ege, real_age	0.462 ± 0.089	0.486
NB	mean_gender, mean_age, mean_emo, sft, ege, real_age	0.460 ± 0.084	0.495
NB	mean_age, mean_emo, sft, ege, real_age	0.459 ± 0.083	0.496
NB	mean_gender, mean_age, sft, ege, real_age, ctdp	0.454 ± 0.085	0.500
NB	lengths, mean_age, sft, ege, real_age, ctdp	0.450 ± 0.074	0.501
NB	mean_age, mean_emo, pft, sft, ege, real_age, ctdp	0.450 ± 0.077	0.502
NB	mean_age, pft, sft, ege, real_age, ctdp	0.446 ± 0.076	0.504
NB	lengths, mean_age, pft, sft, ege, real_age, ctdp	0.438 ± 0.074	0.504
LR	mean_gender, sft, ctdp	0.423 ± 0.047	0.603
LR	sft, ctdp	0.420 ± 0.043	0.612
LR	sft	0.417 ± 0.038	0.635

MMSE Prediction

Figures 40 and 41 illustrate the relationship between the Joint Interpretability Index and the mean RMSE for MMSE prediction across algorithm configurations on the Pareto front. Two curves are shown: one representing the true RMSE values obtained via cross-validation, and another depicting the RMSE values predicted by the symbolic regression model defined in (3.2). The model provides a highly accurate approximation of the observed data, with a RMSE of 0.0036.

$$\text{RMSE} = -2JII + 6.24188 - \frac{0.000495741}{JII - 0.69288} - \frac{1.92156}{JII} \tag{3.2}$$

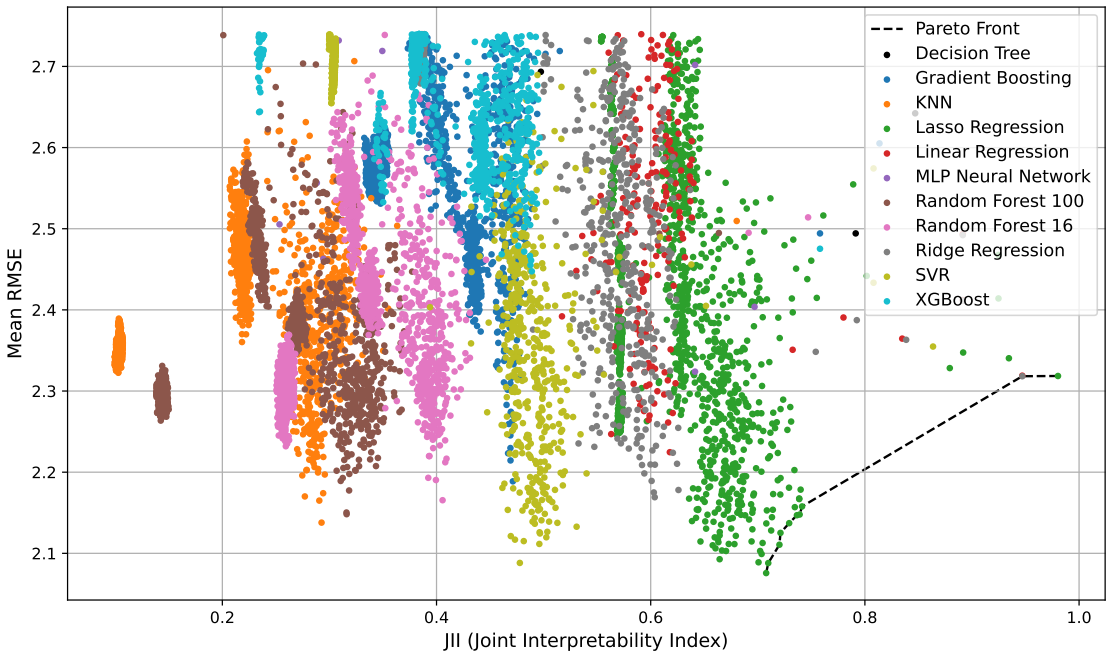

Figure 40. Pareto front: Mean RMSE vs JII.

The trade-off is much more pronounced in the diagnosis task, whereas for MMSE prediction, the best-performing models are also among the most interpretable.

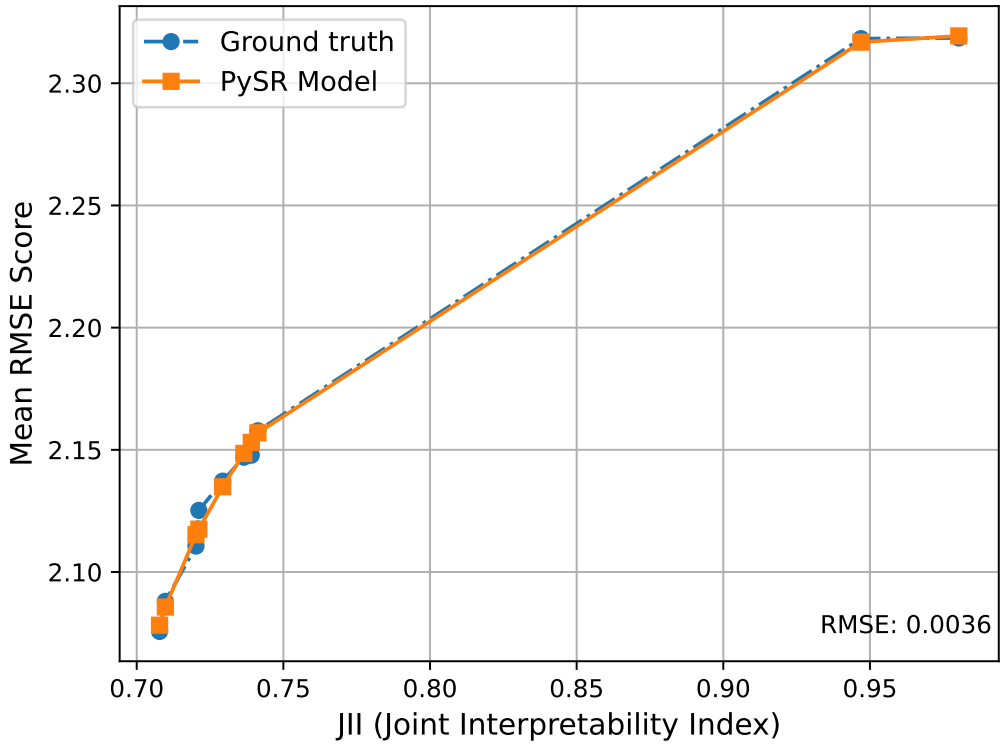

Figure 41. PySR model approximation of Pareto front (MMSE).

Table 23 presents the mean RMSE and corresponding JII values for each model–feature set combination represented on the Pareto front.

Table 23. Pareto optimal models, mean RMSE vs. JII (CV).

Algorithm	Combination	Mean RMSE	JII
Lasso	pft, sft, real_age	2.076 ± 0.327	0.708
Lasso	mean_age, pft, sft, real_age	2.088 ± 0.331	0.710
Lasso	lengths, pft, sft, real_age	2.111 ± 0.359	0.720
Lasso	lengths, mean_age, pft, sft, real_age	2.125 ± 0.366	0.721
Lasso	sft, real_age	2.137 ± 0.317	0.729
Lasso	mean_age, sft, real_age	2.147 ± 0.322	0.737
Lasso	sft, real_age, ctdp	2.148 ± 0.307	0.739
Lasso	mean_age, sft, real_age, ctdp	2.158 ± 0.314	0.741
Ridge	real_age	2.318 ± 0.449	0.947
Lasso	real_age	2.319 ± 0.454	0.980

4. Discussion

The primary goal of this research was to evaluate and compare the performance and interpretability of different machine learning models for detecting cognitive decline through spontaneous speech. The study focussed on traditional machine learning models, which were used to classify cognitive status and predict cognitive assessment scores. This chapter first summarizes the key findings, describes limitations noted, and proposes future directions.

4.1 Key Findings and Insights

This section summarizes the main findings of the thesis, framed in relation to the three research questions introduced in Chapter 1. Each research question is addressed in light of the experimental results and analyses presented in Chapter 3.

RQ1: How do machine learning models, using different combinations of both interpretable features and learned features from pretrained deep learning models, compare in classifying individuals as healthy, MCI, or dementia and predicting cognitive assessment scores?

Diagnosis Classification Results: 11 classification algorithms were evaluated across 2,047 different feature set combinations, yielding a total of 22,517 model–feature combinations for both the development and cross-validation settings. The PROCESS challenge baseline for classification achieved a macro-F1 score of 0.393; any result above this was considered a meaningful improvement. 11,750 combinations (52%) outperformed the official PROCESS challenge baseline on the development set, and 13,619 (60%) outperformed the baseline under cross-validation. These results, summarized in Table 15 of Section 3.2.2, highlight how the top-performing algorithms compared in terms of both their frequency of strong performance and their best F1 scores.

The results show that MLP was consistently among the strongest performers, achieving the best overall cross-validation macro-F1 score of 0.511, and ranking first in the CV results. Gradient Boosting yielded the best single development set performance (0.670), though this did not translate into a similarly strong rank in the CV setting. In fact, despite nearly 70% of feature combinations combined with this model beating the baseline, the best CV score was only 0.445.

Interestingly, simpler models such as Naive Bayes and Logistic Regression also demonstrated robust performance, contributing significantly to the pool of results that beat the baseline. This suggests that when paired with the right feature sets these models can be surprisingly competitive.

The overall trend indicates that both neural and ensemble-based models (e.g., MLP, Random Forest, Gradient Boosting, XGBoost) frequently performed well across a wide range of feature configurations. However, despite beating the baselines with majority of feature set combinations, the cross-validation performance of the ensemble models were only mediocre, and their advantage in performance often comes at the cost of reduced interpretability, an issue explored in more detail later in this chapter.

In contrast, the consistency of models like SVM and Logistic Regression across both dev and CV settings suggests that some degree of robustness can be achieved even with simpler classifiers. This finding is especially relevant for clinical settings, where interpretability and stability across datasets may be more valuable than marginal gains in performance.

MMSE Regression Results: For the regression task of predicting Mini-Mental State Examination scores, 11 regression algorithms were evaluated using the same 2,047 feature set combinations, resulting in 22,517 evaluations for both the development and cross-validation settings. The PROCESS challenge baseline for MMSE prediction achieved a root mean squared error of 2.74, and any model achieving a lower RMSE was considered to outperform the baseline. 15,351 combinations (68%) outperformed the official PROCESS challenge baseline on the development set, and 12,554 (56%) outperformed the baseline under cross-validation. Table 16 summarizes the top-performing algorithms on this task.

In contrast to classification, interpretable models such as Lasso Regression and Ridge Regression performed exceptionally well. Lasso Regression achieved the best overall performance in cross-validation with an RMSE of 2.076, and also ranked first overall. Ridge Regression was similarly competitive, particularly on the development set. Notably, models like K-Nearest Neighbors and Random Forests also performed consistently well across both data splits, beating baselines with nearly 100% of feature set combinations.

One notable trend is the alignment between performance and interpretability: the most successful models in regression were often also the simpler and more transparent algorithms. For instance, Lasso and Ridge Regression, known for their explainability and feature sparsity, outperformed more complex models such as neural networks and XGBoost in cross-validation settings, with MLP beating baselines with only 0.4% of all feature-set combinations.

Interestingly, although Decision Trees achieved the single lowest RMSE (1.658) on the development set, this result did not generalize well to cross-validation, where its performance degraded substantially (2.494 RMSE). This suggests that some of the most striking individual results may not reflect reliable or generalizable performance.

In contrast to the classification task, the regression results point to a more favorable relationship between model transparency and accuracy. Interpretable models were not only competitive, but often superior, highlighting the potential for clinically meaningful cognitive score prediction that does not rely on opaque or black-box architectures.

These findings support a more optimistic outlook for deploying interpretable regression models in remote cognitive health screening, particularly when used to estimate test scores like MMSE from spontaneous speech features.

RQ2: Which acoustic, text-based, and demographic features—either automatically extracted or produced through deep learning—are most indicative of cognitive decline, and how do they contribute to model decisions for both classification and regression tasks?

To determine which features are most indicative of cognitive decline, a detailed analysis was performed across both classification (diagnosis) and regression (MMSE score prediction) tasks. The features considered fell into three broad categories—acoustic, text-based, and demographic—and were further subdivided based on their method of derivation: latent features (e.g., deep learning embeddings such as `xvec`), semantic features (semi-interpretable outputs from deep models, such as `mean_emo` or `ctdp`), and concrete features (directly interpretable, rule-based descriptors such as word counts, pauses, or demographic attributes).

In the classification task focused on diagnosing cognitive impairment, latent acoustic features were the most frequently employed and the most impactful. Notably, the `xvec` embeddings derived from deep neural models were not only widely used but also dominated the top-performing feature combinations. These embeddings may encapsulate complex acoustic information, capturing subtle voice characteristics associated with cognitive decline. The effectiveness of these features is evidenced by the performance of neural network models such as MLPs, which relied heavily on `xvec`, as well as supplementary semantic features like audio-based emotion representations and fluency-related pause features.

In contrast, for the regression task aimed at predicting MMSE scores, the feature land-

scape shifted markedly. Here, concrete and semantic features outperformed latent ones. Semantic and phonemic fluency tasks emerged as the most powerful indicators of cognitive performance. Among these, the number of animals named in a 60-second semantic fluency task (`animals_sft`) proved to be the single strongest predictor of MMSE, aligning with clinical observations that reduced semantic fluency often appears early in cognitive impairment. The number of words beginning with the letter P in the phonemic fluency task (`p_words_pft`) also showed a strong positive correlation with MMSE scores.

Additional predictive power came from acoustic features that reflected voice quality and articulation, such as spectral flux, MFCC coefficients, and shimmer. These were particularly effective in distinguishing between levels of impairment, especially when combined with temporal and prosodic dynamics. For instance, more monotonous or spectrally flat speech (low spectral flux) was commonly associated with more severe cognitive decline. These features contributed to models' ability to differentiate between dementia, MCI, and healthy aging. Interestingly, the `xvec` was less performant in the regression task. However, given that these embeddings were trained for the classification task, and considering the distributional difference of MMSE scores across datasets (see Figure 16), the prediction of MMSE scores based on embeddings indicative of diagnostic labels may be too inaccurate to achieve good RMSE.

Demographic variables, particularly real age, also played a consistent role in the regression models. As expected, increased age correlated negatively with MMSE scores. However, some classification models yielded counterintuitive findings—such as higher predicted age being associated with a reduced likelihood of Alzheimer's dementia—suggesting possible confounds or data distribution imbalances that warrant further investigation.

Overall, the analyses indicate that feature importance is task-specific. Latent acoustic features such as `xvec` are highly effective for classification-based diagnosis, likely due to their ability to capture complex speech patterns in an abstract form. Conversely, in the MMSE regression task, interpretable features derived from fluency tasks and prosodic characteristics not only perform better but also align more naturally with clinical reasoning. This duality highlights the importance of adapting feature selection strategies to the goals of the predictive model—whether it aims for high sensitivity in diagnosis or detailed, interpretable assessment of cognitive scores.

RQ3: What is the trade-off between model performance and interpretability for the diagnosis and cognitive assessment score prediction tasks?

The trade-off between model performance and interpretability was a central theme in

evaluating approaches to both diagnosis and cognitive assessment score prediction tasks. This tension is particularly salient in health-related applications, where high predictive accuracy must often be balanced against the need for transparent, clinically meaningful explanations.

In the diagnosis task, the most accurate models—especially those relying on latent acoustic embeddings such as `xvec`—generally exhibited low interpretability. These features, derived from deep learning models, may capture rich paralinguistic information but offer limited transparency regarding the specific speech characteristics that drive model decisions. For instance, a Multilayer Perceptron using `xvec` embeddings achieved a respectable macro-F1 score of 0.41 in cross-validation. When additional semantic and pause-based features (e.g., `mean_emo`, `ctdp`) were incorporated, this score improved to 0.51, but the underlying feature representations and interactions remained largely opaque. SHAP analyses of these models revealed class-specific patterns of influence, but the interpretability of these patterns remained challenging due to the abstract nature of the latent features.

In contrast, for MMSE score prediction, interpretability and predictive utility appeared more aligned. Concrete features such as word counts from structured fluency tasks, prosodic attributes (e.g., pause durations), and demographic variables not only improved model transparency but also contributed substantially to predictive performance. For example, the Decision Tree Regressor, which uses interpretable thresholds based on features like `animals_sft` (semantic fluency), `spectralFlux_sma3_mean`, and `mfcc3_sma3_mean`, achieved a RMSE of 1.658 on the development set. This model structure allows clinicians to trace individual predictions back to clinically interpretable splits, such as the number of animals named or changes in vocal spectral dynamics.

Further insights were provided by Lasso Regression and SHAP analyses. These revealed that higher scores in semantic and phonemic fluency tasks (e.g., `animals_sft`, `p_words_pft`) and lower real age were positively associated with MMSE, while increased pause counts and use of incorrect or repeated words were negatively associated. Importantly, these relationships are both statistically significant and clinically plausible, further reinforcing the value of interpretable models in regression tasks.

A formal Pareto front analysis using the Joint Interpretability Index reinforced these findings. For classification, models achieving the highest accuracy were typically the least interpretable, illustrating a clear trade-off. However, in the regression setting, the most interpretable model–feature set combinations were also among the best-performing, suggesting a more harmonious relationship between interpretability and utility when

predicting continuous cognitive scores.

While the macro-F1 score of 51% in cross-validation may appear modest, it was achieved under rigorous and systematic evaluation. This value is arguably a more robust indicator of model generalizability than single-split evaluations. Model performance was higher on the development set (67%) and test set (61%). The performance of the challenge organizers' top model improved from 39% on the development set to 55% on the test set. These trends suggest potential distributional differences between splits. As such, the cross-validation result may serve as a conservative estimate of real-world performance. Moreover, the achievable upper bound for classification accuracy in this domain remains unknown. It is plausible that spontaneous speech alone may not encode sufficient signal to support very high performance, and that the 51% result already approaches a meaningful limit under current conditions.

In conclusion, while classification models benefit from complex, latent representations that maximize predictive accuracy, they sacrifice interpretability—a critical concern in clinical decision-making. By contrast, regression models aimed at predicting MMSE scores can leverage interpretable features without compromising performance, offering a promising direction for developing transparent and clinically relevant tools for cognitive monitoring.

4.2 Limitations

Several limitations should be acknowledged in interpreting the results of this study. First, a lack of transparency in the data collection process for the PROCESS dataset introduces uncertainty regarding the temporal relationship between the speech recordings and cognitive assessments. Despite direct inquiries to the dataset authors, it remained unclear whether the MMSE scores and clinical diagnoses were determined before or after the speech recordings. If participants' diagnoses were known beforehand, it is possible that the recording sessions were not diagnostic-naive, weakening the causal link between speech behavior and assessment outcomes. A more robust experimental design would have involved collecting speech first and then administering cognitive assessments, thereby validating the use of speech as a predictive input.

Second, although external datasets were considered for augmentation and comparison, their integration was limited due to distributional mismatches with the PROCESS dataset. These discrepancies—in recording conditions, elicitation tasks, and speaker demographics—undermined their usefulness for direct model training or evaluation. The inclusion of even well-curated external data requires careful alignment, and in this study, their role was necessarily constrained to feature-specific augmentation.

Third, the observed differences between development and test set performance raise concerns about dataset shift and representativeness. The limited number of allowed test submissions and lack of access to test set labels prevented in-depth evaluation, making it impossible to fully analyze or explain these differences.

Finally, it is important to note that the theoretical performance ceiling of speech-based cognitive assessment is unknown. While a macro-F1 of 51% in CV may not suffice for clinical application, it is not clear how much of the cognitive signal is encoded in speech alone. Future work may reveal that such results already approximate the upper limit of what is possible using unimodal speech data.

4.3 Future Work

Several future research directions arise from the limitations and findings of this thesis. First, clearer documentation of dataset collection protocols is essential. In particular, the relationship between speech recordings and cognitive assessments remains ambiguous across current datasets. Future work should prioritize more consistent and purpose-driven data collection designs, where speech is gathered as a direct input to cognitive scoring rather than as an auxiliary measure after scrores and diagnoses are already established.

More general and adaptable modeling strategies could also be explored. Task-agnostic and language-agnostic approaches may help build models that generalize better across different populations, cognitive tasks, and data sources. This direction naturally connects with the integration of additional non-invasive behavioral modalities—such as handwriting, gaze tracking, or digital interaction patterns—which may encode complementary cognitive signals and contribute to more robust and comprehensive screening systems.

Interpretability remains an important consideration, particularly in clinical contexts where it can support trust and decision-making. The success of transparent models in MMSE prediction suggests that future work could aim for outputs that are more aligned with clinical reasoning, ideally developed in collaboration with healthcare professionals. At the same time, the potential and limitations of speech-based cognitive assessment remain uncertain. Rather than relying on standard tests like the MMSE—which draw on a variety of cognitive skills—future research could explore the creation of new, speech-based assessment tasks designed specifically for remote data collection and well-suited to automated analysis.

Together, these directions support a research agenda focused not only on improving performance, but also on enhancing generalizability, interpretability, and clinical relevance.

5. Summary

This thesis explored the use of machine learning to detect cognitive decline from spontaneous speech, focusing on model performance and interpretability. The motivation stemmed from the growing need for scalable, non-invasive tools to aid in detecting conditions like mild cognitive impairment and Alzheimer's dementia—where timely diagnosis is critical for improving patient outcomes. Speech, as a reflection of early signs of cognitive deterioration, serves as a promising modality for such assessments.

The research combined traditional interpretable features—such as fluency and pause metrics—with deep learning-derived acoustic and textual representations. These were used to train machine learning models for two tasks: classifying cognitive status and predicting Mini-Mental State Examination scores. A novel Joint Interpretability Index was introduced to assess the trade-offs between model performance and interpretability, aiming to identify models suitable not just for high accuracy but also for potential clinical deployment.

A comprehensive experimental framework was applied using the PROCESS dataset, supplemented with external corpora for selective augmentation. Extensive cross-validation was performed to evaluate model robustness across a wide range of feature set combinations and model types. The results showed that acoustic features derived from deep learning models provided the greatest benefit for classification, distinguishing between healthy individuals, mild cognitive impairment, and dementia. In contrast, interpretable features—such as semantic fluency, pause frequency, and demographic variables—were more effective in the regression task of predicting Mini-Mental State Examination scores.

Pareto front analysis using the Joint Interpretability Index revealed distinct trends across tasks. In classification, higher accuracy typically required more complex and less interpretable models and feature sets. However, for regression, the best-performing models were also among the most interpretable, suggesting that clinical utility may be more feasible in score prediction tasks. These findings emphasize the importance of considering not only accuracy but also transparency when designing models for sensitive applications like cognitive health screening.

Overall, this study contributes a structured approach to evaluating the dual objectives of performance and interpretability in speech-based cognitive decline detection and provides a foundation for future development of clinically viable remote screening tools.

References

[1] Ziad S. Nasreddine et al. “The Montreal Cognitive Assessment (MoCA): A Brief Screening Tool for Mild Cognitive Impairment”. In: *Journal of the American Geriatrics Society* 53.4 (2005), pp. 695–699. DOI: 10.1111/j.1532-5415.2005.53221.x.

[2] Marshall F. Folstein, Susan E. Folstein, and Paul R. McHugh. “Mini-Mental State: A Practical Method for Grading the Cognitive State of Patients for the Clinician”. In: *Journal of Psychiatric Research* 12.3 (1975), pp. 189–198. DOI: 10.1016/0022-3956(75)90026-6.

[3] World Health Organization. *Ageing and Health*. Last modified October 1, 2024. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health.

[4] S. de la Fuente Garcia, C. W. Ritchie, and S. Luz. “Artificial Intelligence, Speech, and Language Processing Approaches to Monitoring Alzheimer’s Disease: A Systematic Review”. In: *Journal of Alzheimer’s Disease* 78.4 (2020), pp. 1547–1574. DOI: 10.3233/JAD-200888.

[5] Saturnino Luz et al. “Editorial: Alzheimer’s Dementia Recognition through Spontaneous Speech”. In: *Frontiers in Computer Science* 3 (2021), p. 780169. DOI: 10.3389/fcomp.2021.780169.

[6] RNU Voleti, J. Liss, and V. Berisha. “A Review of Automated Speech and Language Features for Assessment of Cognition and Thought Disorders”. In: *IEEE Journal of Selected Topics in Signal Processing* (2019).

[7] Xiaoke Qi et al. “Noninvasive Automatic Detection of Alzheimer’s Disease from Spontaneous Speech: A Review”. In: *Frontiers in Aging Neuroscience* 15 (2023). DOI: 10.3389/fnagi.2023.1224723.

[8] S. Luz et al. “Connected Speech-Based Cognitive Assessment in Chinese and English”. In: *Proceedings of INTERSPEECH 2024*. 2024, pp. 947–951. DOI: 10.21437/Interspeech.2024-1807.

[9] Brian MacWhinney. “Understanding Spoken Language Through TalkBank”. In: *Behavior Research Methods* 51.4 (2019), pp. 1919–1927.

[10] Yin-Long Liu et al. *Clever Hans Effect Found in Automatic Detection of Alzheimer’s Disease through Speech*. 2024. URL: https://arxiv.org/abs/2406.07410.

[11] F. Haider et al. “Detecting Cognitive Decline Using Speech Only: The ADReSSo Challenge”. In: *arXiv preprint* (2021). URL: https://arxiv.org/abs/2104.09356.

[12] Y. Wang et al. “Exploring Linguistic Feature and Model Combination for Speech Recognition-Based Automatic AD Detection”. In: *Proceedings of INTERSPEECH 2022*. 2022, pp. 3328–3332. DOI: 10.21437/Interspeech.2022-723.

[13] Z. Shah et al. “Learning Language and Acoustic Models for Identifying Alzheimer’s Dementia from Speech”. In: *Frontiers in Computer Science* 3 (2021), pp. 624–659. DOI: 10.3389/fcomp.2021.624659.

[14] J. Weiner, C. Herff, and T. Schultz. “Speech-Based Detection of Alzheimer’s Disease in Conversational German”. In: *Proceedings of INTERSPEECH 2016*. 2016, pp. 1938–1942.

[15] L. Hernández-Domínguez et al. “Computer-Based Evaluation of Alzheimer’s Disease and Mild Cognitive Impairment Patients During a Picture Description Task”. In: *Alzheimer’s Dementia (Amsterdam)* 10 (2018), pp. 260–268. DOI: 10.1016/j.dadm.2018.02.004.

[16] E. Edwards et al. “Multiscale System for Alzheimer’s Dementia Recognition Through Spontaneous Speech”. In: *Proceedings of INTERSPEECH 2020*. 2020, pp. 2197–2201.

[17] T. Warnita, N. Inoue, et al. “Detecting Alzheimer’s Disease Using Gated Convolutional Neural Network from Audio Data”. In: *Proceedings of INTERSPEECH 2018*. 2018, pp. 1706–1710.

[18] J. Koo et al. “Exploiting Multi-Modal Features from Pre-Trained Networks for Alzheimer’s Dementia Recognition”. In: *Proceedings of INTERSPEECH 2020*. 2020, pp. 2217–2221.

[19] Y. Pan et al. “Automatic Hierarchical Attention Neural Network for Detecting AD”. In: *Proceedings of INTERSPEECH 2019*. 2019, pp. 4105–4109.

[20] A. Ablimit, K. Scholz, and T. Schultz. “Deep Learning Approaches for Detecting Alzheimer’s Dementia from Conversational Speech of ILSE Study”. In: *Proceedings of INTERSPEECH 2022*. 2022, pp. 3348–3352.

[21] L. Yang et al. “Augmented Adversarial Self-Supervised Learning for Early-Stage Alzheimer’s Speech Detection”. In: *Proceedings of INTERSPEECH 2022*. 2022, pp. 541–545. DOI: 10.21437/Interspeech.2022-943.

[22] A. Balagopalan et al. “Comparing Pre-Trained and Feature-Based Models for Prediction of Alzheimer’s Disease Based on Speech”. In: *Frontiers in Aging Neuroscience* 13 (2021), p. 635945. DOI: 10.3389/fnagi.2021.635945.

[23] M. Rupesh Kumar et al. “Dementia Detection from Speech Using Machine Learning and Deep Learning Architectures”. In: *Sensors* 22.23 (2022), p. 9311. DOI: 10.3390/s22239311.

[24] S. Mirzaei et al. “Two-Stage Feature Selection of Voice Parameters for Early Alzheimer’s Disease Prediction”. In: *IRBM* 39.6 (2018), pp. 430–435.

[25] J. V. Egas López et al. “Assessing Alzheimer’s Disease from Speech Using the i-vector Approach”. In: *International Conference on Speech and Computer*. Springer, 2019, pp. 289–298.

[26] G. Gosztolya et al. “Identifying Mild Cognitive Impairment and Mild Alzheimer’s Disease Based on Spontaneous Speech Using ASR and Linguistic Features”. In: *Computer Speech and Language* 53 (2019), pp. 181–197.

[27] S. Kato et al. “Early Detection of Cognitive Impairment in the Elderly Based on Bayesian Mining Using Speech Prosody and Cerebral Blood Flow Activation”. In: *Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2013, pp. 5813–5816.

[28] Florian Eyben et al. “The Geneva Minimalistic Acoustic Parameter Set GeMAPS for Voice Research and Affective Computing”. In: *IEEE Transactions on Affective Computing* 7.2 (2016), pp. 190–202.

[29] Sophie A. Martin et al. “Interpretable Machine Learning for Dementia: A Systematic Review”. In: *Alzheimer’s & Dementia: The Journal of the Alzheimer’s Association* 19.5 (2023), pp. 2135–2149. DOI: 10.1002/alz.12948.

[30] Wei Ying Tan et al. “A Machine Learning Approach for Early Diagnosis of Cognitive Impairment Using Population-Based Data”. In: *Journal of Alzheimer’s Disease: JAD* 91.1 (2023), pp. 449–461. DOI: 10.3233/JAD-220776.

[31] Scott M. Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017, pp. 4765–4774. URL: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[32] Faiza Iqbal et al. “An Explainable AI Approach to Speech-Based Alzheimer’s Dementia Screening”. In: *SMM24, Workshop on Speech, Music and Mind 2024*. 2024, pp. 11–15. DOI: 10.21437/SMM.2024-3.

[33] PROCESS Challenge. *Prediction and Recognition of Cognitive Decline through Spontaneous Speech (PROCESS) Signal Processing Grand Challenge*. 2024. URL: https://processchallenge.github.io/.

[34] H. Herrmann, T. Kaevand, and L. Anton. *BASE: TalTech's HPC Infrastructure 2020–2024*. TalTech Data Repository. DOI: 10.48726/zf23z-8ba50. Mar. 2025. URL: https://doi.org/10.48726/zf23z-8ba50.

[35] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[36] Wes McKinney. "Data Structures for Statistical Computing in Python". In: *Proceedings of the 9th Python in Science Conference* (2010). Ed. by Stéfan van der Walt and Jarrod Millman, pp. 56–61.

[37] Charles R. Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2.

[38] Scott Lundberg and Su-In Lee. *A Unified Approach to Interpreting Model Predictions*. arXiv:1705.07874. 2017. URL: https://github.com/slundberg/shap.

[39] Tianqi Chen and Carlos Guestrin. *XGBoost: A Scalable Tree Boosting System*. 2016. URL: https://doi.org/10.1145/2939672.2939785.

[40] Plotly Technologies Inc. *Collaborative data science*. 2015. URL: https://plot.ly.

[41] Saturnino Luz et al. "An Overview of the ADReSS-M Signal Processing Grand Challenge on Multilingual Alzheimer's Dementia Recognition Through Spontaneous Speech". In: *IEEE Open Journal of Signal Processing* 5 (2024), pp. 738–749. DOI: 10.1109/OJSP.2024.3378595.

[42] Carole Roth. "Boston Diagnostic Aphasia Examination". In: *Encyclopedia of Clinical Neuropsychology*. Ed. by Jeffrey S. Kreutzer, John DeLuca, and Bruce Caplan. New York, NY: Springer New York, 2011, pp. 428–430. ISBN: 978-0-387-79948-3. DOI: 10.1007/978-0-387-79948-3_868.

[43] Florian Eyben, Martin Wöllmer, and Björn Schuller. "Opensmile: the munich versatile and fast open-source audio feature extractor". In: *Proceedings of the 18th ACM International Conference on Multimedia*. MM '10. Firenze, Italy: Association for Computing Machinery, 2010, pp. 1459–1462. ISBN: 9781605589336. DOI: 10.1145/1873951.1874246.

[44] Johannes Wagner et al. "Dawn of the transformer era in speech emotion recognition: closing the valence gap". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.9 (2023).

[45] Alec Radford et al. "Robust speech recognition via large-scale weak supervision". In: *Proc. ICML*. 2023.

[46] Catarina Botelho et al. “Macro-descriptors for Alzheimer’s Disease Detection Using Large Language Models”. In: *Proceedings of INTERSPEECH 2024*. 2024, pp. 1975–1979. DOI: 10.21437/Interspeech.2024-1255.

[47] Cynthia Rudin. *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*. 2019. URL: https://arxiv.org/abs/1811.10154.

[48] Sven Kruschel et al. *Challenging the Performance-Interpretability Trade-off: An Evaluation of Interpretable Machine Learning Models*. 2024. URL: https://arxiv.org/abs/2409.14429.

[49] Felix Burkhardt et al. “Speech-based Age and Gender Prediction with Transformers”. In: *arXiv preprint arXiv:2306.16962* (2023).

[50] Fuxiang Tao et al. “Early Dementia Detection Using Multiple Spontaneous Speech Prompts: The PROCESS Challenge”. In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2025, pp. 1–2. DOI: 10.1109/ICASSP49660.2025.10889017.

[51] Miles Cranmer. *Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl*. arXiv:2305.01582 [astro-ph, physics:physics]. May 2023. DOI: 10.48550/arXiv.2305.01582.

Appendix 1 – Non-Exclusive License for Reproduction and Publication of a Graduation Thesis[1]

I Erik Illaste

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis "Detection of Cognitive Decline from Spontaneous Speech: Comparing Model Performance and Interpretability", supervised by Tanel Alumäe
   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

12.05.2025

[1]The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.

Appendix 2 - Decision Tree Splits

```
Decision Tree Rgressor with sft pft ege.
RMSE on development set: 1.658.
|--- animals_sft <= 12.50
|   |--- spectralFlux_sma3_mean <= 0.13
|   |   |--- p_words_pft <= 12.50
|   |   |   |--- mfcc3_sma3_mean <= 6.65
|   |   |   |   |--- value: [22.00]
|   |   |   |--- mfcc3_sma3_mean >  6.65
|   |   |   |   |--- shimmerLocaldB_sma3nz_std <= 0.72
|   |   |   |   |   |--- value: [19.00]
|   |   |   |   |--- shimmerLocaldB_sma3nz_std >  0.72
|   |   |   |   |   |--- value: [20.00]
|   |   |--- p_words_pft >  12.50
|   |   |   |--- F2bandwidth_sma3nz_mean <= 1097.53
|   |   |   |   |--- value: [26.00]
|   |   |   |--- F2bandwidth_sma3nz_mean >  1097.53
|   |   |   |   |--- value: [25.00]
|   |--- spectralFlux_sma3_mean >  0.13
|   |   |--- alphaRatio_sma3_std <= 14.77
|   |   |   |--- F1frequency_sma3nz_mean <= 717.60
|   |   |   |   |--- slope500-1500_sma3_mean <= -0.01
|   |   |   |   |   |--- value: [24.00]
|   |   |   |   |--- slope500-1500_sma3_mean >  -0.01
|   |   |   |   |   |--- value: [25.00]
|   |   |   |--- F1frequency_sma3nz_mean >  717.60
|   |   |   |   |--- value: [26.00]
|   |   |--- alphaRatio_sma3_std >  14.77
|   |   |   |--- total_pause_sft <= 21.00
|   |   |   |   |--- F2bandwidth_sma3nz_std <= 448.71
|   |   |   |   |   |--- value: [28.00]
|   |   |   |   |--- F2bandwidth_sma3nz_std >  448.71
|   |   |   |   |   |--- value: [27.00]
|   |   |   |--- total_pause_sft >  21.00
|   |   |   |   |--- value: [29.00]
|--- animals_sft >  12.50
|   |--- spectralFlux_sma3_mean <= 0.10
|   |   |--- F2amplitudeLogRelF0_sma3nz_std <= 82.31
|   |   |   |--- value: [26.00]
```

```
|   |   |--- F2amplitudeLogRelF0_sma3nz_std >  82.31
|   |   |   |--- value: [23.00]
|   |--- spectralFlux_sma3_mean >  0.10
|   |   |--- count_pauses_sft <= 3.50
|   |   |   |--- mfcc2_sma3_mean <= -0.26
|   |   |   |   |--- F1amplitudeLogRelF0_sma3nz_std <= 99.39
|   |   |   |   |   |--- value: [30.00]
|   |   |   |   |--- F1amplitudeLogRelF0_sma3nz_std >  99.39
|   |   |   |   |   |--- value: [29.00]
|   |   |   |--- mfcc2_sma3_mean >  -0.26
|   |   |   |   |--- animals_sft <= 19.00
|   |   |   |   |   |--- F1bandwidth_sma3nz_mean <= 1290.57
|   |   |   |   |   |   |--- value: [27.00]
|   |   |   |   |   |--- F1bandwidth_sma3nz_mean >  1290.57
|   |   |   |   |   |   |--- hammarbergIndex_sma3_mean <= 24.41
|   |   |   |   |   |   |   |--- value: [28.00]
|   |   |   |   |   |   |--- hammarbergIndex_sma3_mean >  24.41
|   |   |   |   |   |   |   |--- value: [29.00]
|   |   |   |   |--- animals_sft >  19.00
|   |   |   |   |   |--- F2amplitudeLogRelF0_sma3nz_std <= 94.84
|   |   |   |   |   |   |--- shimmerLocaldB_sma3nz_std <= 0.78
|   |   |   |   |   |   |   |--- mfcc1_sma3_mean <= 15.35
|   |   |   |   |   |   |   |   |--- value: [29.00]
|   |   |   |   |   |   |   |--- mfcc1_sma3_mean >  15.35
|   |   |   |   |   |   |   |   |--- value: [28.00]
|   |   |   |   |   |   |--- shimmerLocaldB_sma3nz_std >  0.78
|   |   |   |   |   |   |   |--- value: [29.00]
|   |   |   |   |   |--- F2amplitudeLogRelF0_sma3nz_std >  94.84
|   |   |   |   |   |   |--- mfcc4_sma3_std <= 15.84
|   |   |   |   |   |   |   |--- value: [27.00]
|   |   |   |   |   |   |--- mfcc4_sma3_std >  15.84
|   |   |   |   |   |   |   |--- value: [28.00]
|   |   |--- count_pauses_sft >  3.50
|   |   |   |--- F3amplitudeLogRelF0_sma3nz_mean <= -136.52
|   |   |   |   |--- value: [25.00]
|   |   |   |--- F3amplitudeLogRelF0_sma3nz_mean >  -136.52
|   |   |   |   |--- count_pauses_sft <= 8.50
|   |   |   |   |   |--- value: [28.00]
|   |   |   |   |--- count_pauses_sft >  8.50
|   |   |   |   |   |--- value: [27.00]
```