Universiteit Leiden

Master Computer Science

[Differentiating Ordered from Disordered Language In Older Adults Using the Cookie Theft Picture Description Task: A Multimodal Machine Learning Approach]

Name:	[Belal Mian]
Student ID:	[s3737896]
Date:	[11/08/2025]
Specialisation:	[Data Science]
1st supervisor:	[Prof.dr. M.R. Spruit]
2nd supervisor:	[Dr. B.M.A. van Dijk MSc]

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1

Contents

Differentiating Ordered from Disordered Language In Older Individuals Using the Cookie Theft Picture Description Task: A Multimodal Machine Learning Approach

Belal Mian

s3737896

ABSTRACT

This study investigates the application of three machine learning models(Random Forest, SVM, XGBoost) and a baseline linear regression model to distinguish disordered language patterns associated with Alzheimer's Disease (AD) from typical language use in the older individual populations, addressing the research question: *"How can we use machine learning to distinguish normal from disordered language use in monitoring Alzheimer's Disease in older individuals?"*

This study focuses on Alzheimer's Disease, leveraging linguistic and auditory patterns as a diagnostic tool. The dataset for this study comprises 674 participants(337 AD, 337 control) taking the Cookie Theft picture description test, sourced from the DementiaBank database. Textual features were extracted using Empath and sentence embeddings, while audio features were obtained from the emobase feature set from OpenSMILE.

The feature selection consisted of three steps: statistical significance testing $(p < 0.05)$, effect size filtering $(Cohen's\ d \geq 0.5)$, and multicollinearity reduction$(Pearson's\ r > 0.8)$. All models achieved ROC AUC scores of 0.73–0.76 and recall scores of 0.72–0.73, achieving results comparable to those reported in the literature. In contrast to expectation, audio features contributed comparably to text features. The performance of the machine learning models compared to each other and to the baseline linear regression model was similar, suggesting feature quality outweighed model choice. Despite promising results, limitations include dataset size, task specificity, and non-specialized feature tools. Future work should expand datasets, tailor features specifically to detect AD, and refine multimodal approaches by employing a hybrid approach. This approach leverages the strengths of separate models that can be used to train on text and audio features independently, combining their outputs via a secondary model to improve overall prediction performance.

Keywords: Machine learning, Disordered language, Alzheimer's Disease, Cookie Theft task, Multimodal analysis, Feature selection

1 INTRODUCTION

Mental health problems are the leading cause of disability[52]. This has significant economic consequences, costing the U.S. economy $282 billion annually[67]. Given this substantial impact, further research into mental health conditions is crucial, not only to prevent prolonged periods of disability, but also to develop better interventions that can shorten them. A particularly vulnerable segment of the population are older individuals(65 years and older), who face unique challenges in mental health assessment. By 2030, one in six people worldwide will be aged 60 years or older[53]. The prevalence of age-related mental health conditions such as Alzheimer's Disease (AD) is expected to increase significantly as the global population ages. AD is one of the most prevalent mental health disorders among older adults, making it a critical focus area to improve diagnostic and monitoring methods[51]. The number of AD patients is expected to reach 78 million in 2030 and 139 million in 2050, making early detection of it more urgent[1]. Cognitive decline through normal aging or neurodegenerative conditions such as AD complicate traditional diagnostic methods that often rely on self-reporting. These conventional approaches may be less effective for older adults experiencing communication difficulties or cognitive impairments. One promising area of research involves using language markers to improve the diagnosis of mental health conditions such as AD. These markers include the words people use or the topics they frequently discuss. Research has shown that individuals with mental health disorders exhibit distinctive linguistic characteristics, such as increased use of negative words, disorganized sentence structures, or reduced speech fluency[42]. Identifying these markers, however, requires analyzing large amounts of linguistic data, often containing subtle and complex patterns that may not be evident. In light of these obstacles, different methods are required to diagnose mental health conditions in older adults, especially for those struggling with communication. Machine learning may offer a solution. By analyzing subtle written and/or spoken linguistic patterns that may go undetected otherwise, it can help identify mental health concerns, even when communication is impaired[36]. This approach could be valuable for diagnosing or monitoring mental health conditions such as Alzheimer's Disease in older adults. This reduces the reliance on self-reporting while improving diagnostic accuracy. With this in mind, an important research question arises:

"How can we use machine learning to distinguish normal from disordered language use in monitoring Alzheimer's Disease in older individuals?"

In this study, disordered language will be defined as difficulties in using and understanding spoken or written language. This term aligns with the definition of language disorders used by the American Speech-Language-Hearing Association. They define a language disorder as an impaired comprehension and/or use of spoken, written and/or other symbol systems[8][6]. This can include challenges in speaking, listening, reading, or writing, making it difficult for individuals to express their thoughts or comprehend others[17]. To answer the research question, this study proceeds as follows: Chapter 2 establishes a theoretical framework where relevant literature on Alzheimer's Disease progression is reviewed. This also contains linguistic markers of cognitive decline. This is followed by the challenges in older adults' mental health research and current gaps that motivate this study. Chapter 3 details the methodology, including data collection from the DementiaBank database using the Cookie Theft picture description task, participant demographics. Afterwards, the feature extraction methods, statistical analysis, feature selection, the used machine learning models, and the training process will be discussed. Chapter 4 presents the results. This includes feature selection effectiveness, assessing model performance based on performance metrics, confusion matrix analysis, hyperparameter tuning value distribution, and feature importance across models. Chapter 5 discusses these findings, highlighting unexpected insights and limitations, while Chapter 6 concludes with a summary of key contributions, findings, and future directions.

2 LITERATURE REVIEW

This literature review begins by examining the theoretical foundations of language patterns in Alzheimer's Disease (AD), and continues by exploring challenges in mental health research among older adults. It then identifies current research trends and gaps that this study addresses.

2.1 Theoretical Framework

AD significantly affects the population of older adults. AD encompasses three stages: preclinical, prodromal, and dementia AD [24]. In the preclinical stage, changes in the brain occur without noticeable symptoms[24]. Prodromal AD, often called Mild Cognitive Impairment(MCI), represents the early stages of cognitive decline. This is noticeable, but not yet severe [24]. Research has shown that individuals in the early stages of AD exhibit reduced syntactic complexity, meaning that they were more prone to use simpler sentences[12]. These sentences often lack relative clauses (dependent clauses that modify nouns, e.g., "who," "which," or "that"), and contain fewer multiple verbs or noun phrases[12]. Additionally, the proportion of grammatically correct sentences also tends to decrease in the early stages of Alzheimer's Disease [12]. Another observed pattern is an increased reliance on using pronouns and verbs, accompanied by a reduced noun usage[12]. For example, a patient might say '*He is doing that.*' instead of '*The man is opening the door.*'. Dementia AD is the final stage, where cognitive decline becomes severe[24]. Collectively, these three stages affect approximately 416 million people, representing 22% of the global population aged 50 and above[24].

Given the high prevalence of AD among older individuals, diagnosing it early becomes increasingly crucial. Early intervention may slow the progression of the disease and could preserve some neural structures[33]. This may potentially delay the emergence of certain symptoms in patients[33]. One promising method for diagnosing conditions such as AD is through the analysis of language, both verbal and written. Language plays a fundamental role in shaping cognition and perception, making it a powerful indicator of an individual's psychological state[68]. Research suggests that linguistic features can offer critical insights into mental health conditions, with specific language patterns serving as markers of underlying psychological states[45]. An individual's psychological state often manifests in their communication style, providing valuable cues for diagnosis. For example, individuals with mental health difficulties may use distinctive language patterns, such as an overuse of singular personal pronouns (e.g., "I", "me") and negative emotion words, compared to those without such difficulties[42].

AD, particularly in its early stages, also influences speech. A study found that a decline in speech rate could be an early indicator of AD[68]. This decline in speech rate may reflect broader cognitive and neurological changes occurring in the brain[68]. Furthermore, AD patients exhibit an unstable pitch, where their pitch can change frequently[44]. Their average pitch also tends to be lower compared to cognitively healthy individuals[44]. Another characteristic feature of AD is fluctuating speech volume, which can vary irregularly throughout conversation[44]. This likely reflects reduced motor control or impaired self-monitoring[44]. Beyond speech acoustics, patients with AD often display a reduced or absent grammatical structure in their speech, along with a higher frequency of repetitiveness and misspellings in their writing[20]. The use of generic terms to refer to subjects, such as "boy", "girl", or "woman" instead of more specific terms like "son", "brother", "sister", or "mother", is also associated with a higher risk of Alzheimer's Disease[20]. In contrast, mentioning more specific details, such as "dishcloth" and "dishes", may be linked to a lower risk of AD[20]. Given these findings, linguistic markers may serve as a valuable tool in diagnosing AD. In some cases, diagnoses based on linguistic markers have outperformed those made through traditional clinical methods[69]. This highlights the importance of considering linguistic markers when diagnosing mental health disorders such as AD in the older adult population, providing a non-invasive and potentially more accurate method of early diagnosis and intervention. The linguistic markers for Alzheimer's Disease are summarized in table 1.

Alzheimer's Disease Linguistic Marker	Meaning
Reduced syntactic complexity	Simpler sentences that lack relative clauses such as "who" or "which" and contain fewer multiple verbs or noun phrases.
The proportion of grammatically correct sentences decreases	Fewer sentences made by AD patients are grammatically correct.
Increased reliance on using pronouns and verbs, accompanied by a reduced noun usage	Instead of saying "A woman is watching TV, eating dinner while a man watches.", they say "She is sitting and eating something. He looks."
Declining speech rate	Speaking slower.
Along with a higher frequency of repetitiveness and misspellings in their writing	Addressing subjects multiple times and misspelling words.
The use of generic terms to refer to subjects	Saying "boy", "girl", instead of more specific terms like "brother" or "sister".

Table 1. Linguistic Markers for Alzheimer's Disease

2.2 Challenges Regarding Research on Alzheimer's Disease of the Older Individuals

While the use of linguistic markers has been established as an effective tool for diagnosing AD, especially in older individuals, requires analyzing large amounts of linguistic data, which presents unique challenges. Older adults are less likely to participate in mental health studies such as those for AD [47]. Factors contributing to this include decreased communicative abilities or cognitive impairments caused by conditions such as AD[47]. This can lead to disorderly language where speech becomes incoherent, fragmented, or difficult to follow[47]. An example of disordered language is Alogia, which can be found in patients with AD[42]. Alogia is characterized by a noticeable decrease in quality and quantity of speech[42]. Individuals with Alogia often do not speak until spoken to, give short answers, pause for a time between words and/or sentences and exhibit a flat vocal tone[42].

The second challenge regarding studying AD in older individuals is their health. Those challenges include general frailty and mobility issues, both of which are exacerbated by AD[14]. Other challenges involve sensory deficits that frequently disqualify individuals from clinical trials and broader participation. Examples of sensory deficits are loss of hearing and vision. Both of which are linked to an increased risk of AD[35]. As previously mentioned, a large amount of linguistic data needs to be analyzed to identify linguistic patterns associated with Alzheimer's Disease. One potential solution is the use of machine learning models. They excel in analyzing large amounts of data, allowing for the capture of subtle linguistic patterns that otherwise may have gone unnoticed. This makes machine learning models well-suited for this task and some have already been successfully employed in some studies.

2.3 Current Research and Research Gaps

While research on linguistic markers for various mental health disorders is well-established, there has been relatively little focus on the older individuals' population[47]. Most recent studies have focused on general populations, leaving a gap in research specifically targeted towards older adults. One example of this is Spruit et al.(2022). They conducted a study on linguistic markers for mental health disorders, using multiple machine learning and deep learning models, to predict the presence of a mental disorder and, if present, which one. For both tests, the study employed five models. Those models included the Decision Tree, Random Forest, Support Vector Machine(SVM), fastText and RobBERT. The features for the Decision Tree, Random Forest and SVM were both extracted with spaCy and LIWC. The dataset in the study was sourced from the Verhalenbank ("Storybank") of the University Medical Centre Utrecht (UMCU) in The Netherlands, where stories related to mental illness were collected from those who have or had psychiatric issues or who were in contact with people with psychiatric issues[64]. While Spruit et al. (2022) did make significant strides in identifying linguistic markers in the Dutch language for mental health disorders, the study lacks the focus on any specific age group. This study will specifically focus on older adults. Furthermore, this study specifically focuses on Alzheimer's Disease. Moreover, this study will also collect audio data, whereas Spruit et al.(2022) focused solely on textual data. The combination of textual and audio data may bring more insight into the language patterns that are associated with Alzheimer's Disease than possible if only one of the two were to be used.

Another study saw the application of unimodal and multimodal approaches to predict the presence of a mental disorder[17]. In this study, a unimodal approach means analyzing one type of data, either text or audio, whereas a multimodal approach analyzes both[17]. The study contains two separate unimodal approaches, one for text and one for audio, and one multimodal approach that combined text and audio data. The goal of this study was to compare the accuracies of those models and to emphasize the importance of multimodal integration in the field of mental health diagnostics and to set the stage for future research[17]. While this study did contribute in analyzing multiple types of data, it also lacks the focus on any specific age group and the focus on AD specifically. Given that cognitive impairments can significantly impact both speech and text data, the analysis of the combination of both types of data may offer unique insights into recognizing the linguistic patterns associated with AD in older individuals.

3 METHODOLOGY

Building upon the theoretical framework and research gaps identified in the literature review, this section outlines a multimodal approach to address the research question. The multimodal approach leverages both textual and audio features derived from a standardized cognitive assessment task, the Cookie Theft picture description task. This section first describes the data collection process followed by the demographics of the participants. It then explains the preprocessing steps involving statistical feature selection, and a discussion of the application of three machine learning models with optimized hyperparameters and a baseline linear regression model.

3.1 Data Collection

The data used in this study was collected from three studies in the DementiaBank database. All three studies used the Cookie Theft description test. This is a test where participants have to describe as many of the events that occur in the picture as possible[14]. The Cookie Theft picture can be found in figure 1. The picture illustrates a mother drying the dishes next to the sink in the kitchen. The sink is overflowing

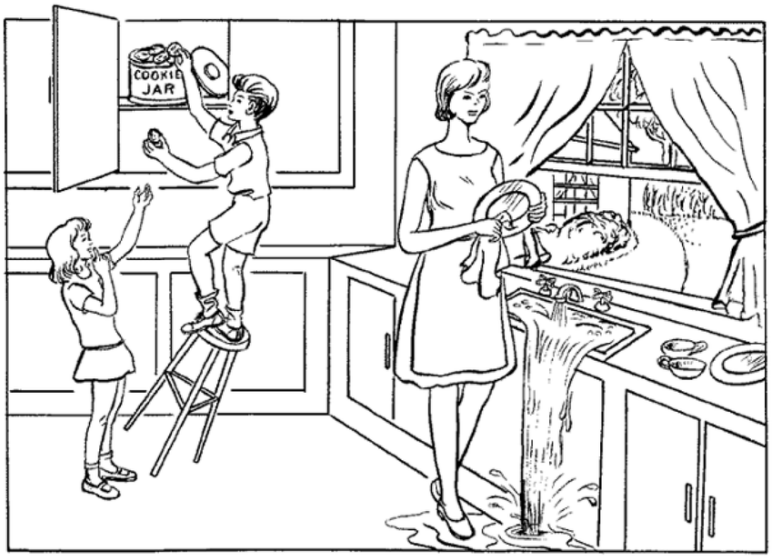

Figure 1. The Cookie Theft Picture

with water due to the mother leaving the tap open and not paying attention. At the same time, two children, a boy and a girl are attempting to take cookies from a jar when their mother is not looking. The boy uses a stool to get up to the cabinet, where the cookie jar is located. The stool is unstable. The girl stands next to the stool and has her hands stretched to receive cookies[14]. Participants were sometimes pointed to neglected features of the picture by the researcher and were asked to elaborate if the response of the participant was less than might be expected given his or her apparent potential[14]. This test is used to evaluate language and cognitive functions[14]. More specifically, it is utilized to prompt various aspects of linguistic abilities of the patients such as spontaneous speech, vocabulary usage, grammatical structures, and the ability to organize and convey information coherently[14].

As previously mentioned, three studies were selected for data collection. Not all samples from the third study were included for analysis. In this particular dataset, the control and AD groups were not specified explicitly. However, this could be partially inferred based on certain indicators, such as having had a stroke, difficulty in thinking, having memory troubles and having had a diagnosed mental health condition(either past or present). Only samples without these indicators were included and classified as part of the control group. In all three studies, older individuals with and without AD were asked to describe the Cookie Theft picture. These descriptions were recorded and transcribed. The transcripts included basic demographic information such as age, sex, and diagnosis (dementia, Alzheimer's disease, MCI, vascular dementia, control)[41][9]. To further illustrate this, a subset of samples is shown in table 2. The audio and the transcripts were used as input for this study.

Language	Researcher	Role	Age	Gender	Diagnosis
eng	Pitt	Participant(PAR)	57	male	ProbableAD
eng	Pitt	Participant(PAR)	76	female	MCI
eng	Pitt	Participant(PAR)	56	male	ProbableAD
eng	Pitt	Participant(PAR)	53	male	ProbableAD
eng	Pitt	Participant(PAR)	75	female	MCI
eng	Pitt	Participant(PAR)	55	male	ProbableAD
eng	Pitt	Participant(PAR)	77	female	MCI
eng	Pitt	Participant(PAR)	59	male	ProbableAD

Table 2. Subset of Samples

3.2 Participant Demographics

This section describes the demographics of the participants. This can be found in table 3. The total number of participants included in this study is 674. Males accounted for 251(37.2%) whilst females accounted for and 423 (62.8%). The overall mean age was 69.47 years with a standard deviation of 9.33, with ages ranging from 46 to 99 years. Male participants had a mean age of 68.69 years with a standard deviation of 9, ranging from 50 to 96 years. Female participants had a mean age of 69.94 with a 9.51 standard deviation, with an age range of 46 to 99. The AD group consisted of 337 participants (50% of the total sample), which included 142 males (42.1%) and 195 females (57.9%). The mean age in this group was 72.05 (SD = 8.75), with an age range of 47 to 91 years. The mean age for male participants in the AD group was 70.66 years (SD = 8.76), while females had a mean age of 72.94 years (SD = 8.65). The control group comprised 337 participants (50% of the total sample), where 117 participants were males (34.7%) and 220 were females (65.3%). The mean age of the control group was 66.26 years (SD = 9.04), with an age range of 46 to 99 years. Male participants had a mean age of 66.04 years (SD = 8.66), and females had a mean age of 66.38 years (SD = 9.27) for females.

Group	Total (n)	Males (n, %)	Females %(n, %)	Mean Age (Years)	Age Range	Mean Male %Age	Mean Female Age
All Participants	674	251 (37.2%)	423 (62.8%)	69.47 ± 9.33	46-99	68.69 ± 9.00 (50–96)	69.94 ± 9.51 (46–99)
AD Group	337 (50.0%)	142 (42.1%)	195 (57.9%)	72.05 ± 8.75	47–91	70.66 ± 8.76 (50–90)	72.94 ± 8.65 (47–91)
Control Group	337 (50.0%)	117 (34.7%)	220 (65.3%)	66.26 ± 9.04	46–99	66.04 ± 8.66 (50–96)	66.38 ± 9.27 (46–99)

Table 3. Demographics of Participants by Group

3.3 Preprocessing

Preprocessing the features is a critical step to reduce the noise and improve the quality of the initially extracted features. This section describes the systematic process of feature cleaning, statistical analysis, and selection applied to the textual and audio features to identify the most discriminative patterns associated with disordered language in older individuals. Feature selection reduces computational complexity and improves classification accuracy due to the reduction of the noise.

3.3.1 Feature Extraction

As mentioned earlier, the data used for the experiments were the audio files and their transcriptions. To extract features from the transcriptions, a modified Empath-based approach was used. Empath is a tool that categorizes words in a given text[21]. Empath leverages word embeddings to identify words that are semantically similar to words in its dictionary in each category[21]. This enables the tool to categorize words that do not appear in its dictionary[21]. It will categorize the words ‘bleed’ and ‘punch’ into the violence category[21]. Empath uses a neural embedding trained on over 1.8 billion words of fiction to identify connotations between words and phrases[21]. This allows for the creation of new lexical categories on demand[21]. The drawback of this method is that words in a text are assigned a category solely based on cosine similarity within a fixed vector space, meaning that each word has one vector representation, even if it has multiple definitions[21]. An example of this is the word ’bank’, which could refer to the financial institution or a river bank depending on the context. Instead of relying on Empath’s default keyword matching, each sentence in the transcript was embedded using the sentence transformer ’all-mpnet-base-v2’. Then, the cosine similarity was computed between the sentence embedding and the embeddings of the predefined Empath category labels. Each sentence was represented as a vector of similarity scores, indicating the degree to which it aligned with multiple Empath categories. This allowed for more flexible and context-sensitive categorization. This may prove useful as patients with Alzheimer’s Disease can exhibit unique linguistic patterns, such as repetitiveness, reduced grammatical structure, and using general terms such as boy or girl instead of brother or sister, that are harder to capture

with predefined lexical categories[20].

The audio features were extracted using OpenSMILE. OpenSMILE is a toolkit used for extracting audio features[19]. It also classifies speech and music signals[19]. The 'emobase' feature set was used to collect the audio features for this study. Emobase is a feature set that can capture features such as speech rate, pitch of the voice, jitter(sudden change in pitch) and shimmer(sudden change in loudness). These are all audio indicators for AD[39]. Emobase is an extensive feature set with 988 features and used several times for cognitive monitoring[3][48][32]. Besides the extensive feature set, another advantage of OpenSMILE is that it is written in c++, which makes it computationally efficient and suitable for real-time audio processing. Real-time audio processing means that the audio is analyzed as it is generated, which makes identifying trends faster[15]. An overview of all the features can be found in table 4.

Feature Set	# Features	Description
Empath (textual)	194	Semantic and emotional categories (e.g., sadness, vacation, health) derived from linguistic analysis using neural embeddings [21].
OpenSMILE emobase (audio)	988	Acoustic low-level descriptors (LLDs) such as pitch, MFCCs, intensity, and jitter, combined with 21 statistical functionals (e.g., mean, std, skewness) [7]

Table 4. Overview of the Textual and Audio Features Used in the Study.

3.3.2 Statistical Analysis

The statistical analysis was conducted as part of the preprocessing of the features. The goal of the statistical analysis was to determine the features that were the most significant when distinguishing the speech of individuals with Alzheimer's Disease and that of healthy controls. This provides a baseline and enhances the understanding of the inherent differences between AD patients and control subjects[26]. Each step of the statistical analysis was performed on all features. The statistical analysis began by calculating descriptive statistics such as the mean, variance, and frequencies of all features for both the AD patients and the control group. These descriptive statistics served as a foundation that facilitated an initial understanding of the differences between the groups and helped contextualize outcomes observed in later stages of the analysis.

The next step in the statistical analysis was the calculation of Cohen's d, which measures the difference between the means of the textual and audio features and expresses that difference in terms of the standard deviation. This is called the effect size, which assesses the magnitude of the observed differences. Afterwards, the Welch t-test was performed to determine the p-values of the features. The Welch t-test was chosen over the standard Student's t-test because it does not assume equal variances between groups, making it more appropriate given the heterogeneity in linguistic and acoustic features between participants with and without AD. This test allowed for the obtainment of a p-value for each feature, indicating the likelihood that the observed difference occurred by chance.
Empath has a total of 194 categories and the emobase feature set of OpenSMILE has 988 features, totaling 1182 features[21][7]. The number of false positives with a p-value of 0.05 would be approximately 59. The number of false positives increases with the total number of features. This is called the multiple testing problem[2]. This was addressed by applying the Holm-Bonferroni correction, which adjusts the significance threshold(p-value) based on the number of tests[31]. The final step involved visualization. An effect size distribution plot is included to illustrate the magnitude and distribution of the differences in audio and textual features between AD patients and the control group. No other visualizations are presented, as the audio features were difficult to interpret directly. Therefore, the effect size distribution plot remains the only viable option to illustrate the differences between the audio and textual features.

3.3.3 Feature Selection

Before preprocessing, the dataset was labeled programmatically, where '1' was used for AD patients and '0' for the control group. During preprocessing, missing values (NaNs) were checked across all features. For features that contained missing values, values were imputed using the mean of the respective feature column. The selection began with only retaining the features with a p-value of less than 0.05, based on Welch's t-test results. Afterwards, from the remaining features, those with an effect size below 0.5 were excluded from further analysis. This ensured that only features with a moderate discriminatory effect were retained[13]. Prior to the final step of the feature selection process takes place, a correlation matrix was made to assess multicollinearity. Features that had a Pearson correlation coefficient greater than 0.8 were considered highly correlated. In such cases, the feature with the larger effect size was selected. In the case of an equal effect size, the feature with the higher variance was selected. If both the effect size and variance were identical, the tie was broken by selecting the feature with the name that appears first in

alphabetical order. A second correlation matrix was generated to verify the reduction in redundancy of the features. The correlation threshold value was chosen due to it being a common value chosen in the literature[65]. While a higher threshold value such as 0.9 would be more lenient and retain more features, it would also risk not reducing multicollinearity sufficiently. The chosen value was selected due to it striking a balance between retaining features and the reduction of multicollinearity.

3.4 Models Used

To discern disordered language from ordered language, several machine learning models were used. One of which was the Random Forest model. A Random Forest model is an ensemble method that works by constructing several decision trees[10]. Each tree is trained on a random subset of the samples with replacement[10]. With replacement means that certain samples may appear more than once, while others may not appear at all[10]. Moreover, only a subset of predictor variables or features is considered at each split during the tree-building process. The Random Forest model mitigates overfitting using a combination of bootstrapping and random feature/predictor selection and captures complex, nonlinear relationships that may be present in the data[40]. Hason and Krishnan (2022) used a Random Forest model to analyze 157 audio files of participants taking the Cookie Theft picture description task and achieved an accuracy of 82.2%. This finding suggests that the Random Forest model is suited for this task[27]. Moreover, Random Forest models offer several advantages such as robustness to overfitting and the ability to handle highly non-linear data[59]. Furthermore, it can provide interpretable results when detecting for Alzheimer's[4]. Additionally, you can make use of parallel processing to speed up the process of training the model. Parallel processing would divide training each tree over each available core or thread of the CPU.

The second machine learning model that is used is the Support Vector Machine(SVM). Support Vector Machines work by finding the hyperplane that best separates data points of different classes in a high-dimensional space[28]. The best separation means that the distance between the hyperplane and the nearest data points from either class, known as support vectors, is as large as possible[28]. The maximization of the margin between the hyperplane and the data points increase the model's ability to generalize to unseen data[28]. SVMs are able to capture nonlinear relationships by using the kernel trick[30]. The kernels map the data to a higher dimensional space where the data is able to be linearly separated[30]. SVMs have been successfully used in the past for AD classification[63]. One study even achieved a 94.5% accuracy score for detecting distinguishing between patients Alzheimer's and the healthy control group[43]. The SVM is also capable of capturing non-linear relations through the different kernels, which is important when you are working with more than 2 features[30]. Another added advantage to the usage of the Support Vector Machine is that they remain computationally efficient even when the number of features exceed the number of samples, which is the case for this study[55].

The third machine learning model to be used is the XGBoost model. The XGBoost model is an ensemble technique that combines multiple weak learners to enhance the prediction capability of the model[49]. The weak learners that are used are often decision trees[49]. Each new tree is trained on the residual errors made by the current ensemble[49]. This is done to gradually reduce the prediction error of the model[49]. Similar to the previous two models, XGBoost has also been used in medical cases concerning the detection of Alzheimer's when used on data of the Cookie Theft picture description task. One study achieved an accuracy of 78% with the XGBoost model, highlighting its potential[38]. Additionally, XGBoost is able to capture complex, non-linear relationships within the data, making it suitable for complex medical cases like the one discussed in this study[5]. Moreover, it can capture feature interactions without explicitly modeling them[70]. This is advantageous because it simplifies the modeling process.

The fourth and final model used is the linear regression model, which served as a baseline to evaluate whether the added complexity of the other machine learning models translate into meaningful performance gains and can provide value beyond what a simple linear classifier achieves. All models were trained on the same preprocessed features and evaluated using the same metrics.

3.5 Training Process

The dataset used in this study consists of both audio and textual data. The features were extracted using Empath, where categories are assigned based on cosine similarity between the sentence embeddings and predefined category terms. Audio features were derived from the emobase feature set provided by OpenSMILE. The classification model was trained on these features to distinguish patients with AD from the control subjects. As described earlier, the dataset was preprocessed by removing features that are statistically irrelevant and had an effect size smaller than 0.5. To mitigate multicollinearity, features were filtered based on pairwise correlation. In cases of a high correlation, only the feature with the highest effect size was retained. If those were equal, the feature with the highest variance was selected. Basic class distribution statistics were calculated to identify potential class imbalances. The hyperparameters for the final model are obtained by nested cross-validation, which helps prevent overfitting[22]. The dataset is divided into 5 folds. In each iteration, one fold is held out as the test set and the remaining folds are used for training.

The nested cross-validation features an inner cross-validation loop that performs grid search over key hyperparameters (e.g., tree depth, number of estimators), while the outer loop evaluates the performance of the model with the hyperparameters obtained by the inner loop on the test set. The hyperparameters leading to the highest performance across folds are chosen for the final model. To ensure that the results were not due to chance, the evaluation metrics were averaged across 10 data partitioning states, generated by varying the random seed from an initial value of 42. The hyperparameters that were tested were multiples of their default values, as this had been shown to be an effective approach for hyperparameter tuning[56]. The hyperparameters for each model that were trained and their values will be shown in Tables 2, 3, and 4, with their default values in bold.

Random forest parameters	**Hyperparameter values**		
n_estimators	50	**100**	200
max_depth	**None**	10	30
min_samples_split	**2**	4	8
min_samples_leaf	**1**	2	4

Table 5. Random forest hyperparameters and their values

Support Vector Machines parameters	**Hyperparameter values**		
C	0.1	0.5	**1**
Kernel	linear	poly	**rbf**
Degree	2	**3**	6
Gamma	**scale**	auto	0.1 (float $\geq$ 0)

Table 6. Support Vector Machines hyperparameters and their values

XGBoost parameters	**Hyperparameter values**		
Learning Rate	0.01	0.1	**0.3**
max_depth	3	**6**	9
Gamma	**0**	0.1	0.2
Subsample	0.5	0.75	**1**

Table 7. XGBoost hyperparameters and their values

The final model is trained on 80% of the dataset and tested on the remaining 20%, as that is the most commonly used split[18]. Its performance is evaluated using several metrics. Given the critical nature of diagnosing Alzheimer's, the performance metrics chosen assess the model's ability to accurately diagnose Alzheimer's cases while accounting for misclassifications. The accuracy evaluates the model's ability to identify cases correctly. Precision quantifies how many predicted positive cases are actually correct and thereby accounting for false positives. Recall measures how many actual positive cases are correctly identified, thus accounting for false negatives. The F1-score is the harmonic mean between precision and recall. The harmonic mean is calculated as the number of values divided by the sum of their reciprocals. The last performance metric is the ROC AUC, which evaluates how well the model is able to distinguish between the two classes across various classification thresholds. Additionally, feature importance scores are extracted to highlight the most discriminative features, from which modality-specific contributions can be analyzed to understand the relative importance of audio and textual inputs. To ensure variability

and robustness in the evaluation, the splitting of the dataset into training and testing subsets, as well as the creation of folds in the nested cross-validation process, were performed using multiple random seeds. This approach enables the generation of different training and testing partitions, as well as varied fold compositions across runs. A total of 10 random seeds were employed, and the reported results represent the average values of the relevant performance metrics and figures computed across all runs.

4 RESULTS

This section presents the findings of the analysis on distinguishing disordered language patterns in older individuals using machine learning approaches. It begins by detailing the feature selection process that served as a foundation for further analysis, followed by an evaluation of model performance across multiple metrics. Subsequent sections explore class-wise prediction patterns through confusion matrix analysis, examine hyperparameter value preferences that emerged during tuning, and investigate feature importance to understand how different modalities contribute to classification decisions.

4.1 Feature Selection via Statistical Testing and Correlation Analysis

To improve the interpretability of the model and reduce the dimensionality of the feature space, statistical and correlational filtering was applied. This process involved three stages: significance testing, effect size filtering, and correlation thresholding. The feature set consists of 1182 features, of which 635 were statistically significant. Figure 2 shows the distribution of effect sizes across all features. Effect sizes represent the difference in a given feature between the AD and control groups, whereas the density indicates the number of features that have a given effect size. A positive effect size means that the AD group scores higher on those features than the control group and vice versa for a negative effect size. Values around 0.2 indicate small effects, 0.5 medium effects, and 0.8 or higher large effects[13]. The peaks for both modalities show that the majority of the features have a small effect in the distinction between the two groups. This illustrates that only a small subset of features exhibit meaningful differences between the AD patients and the control group, underscoring the need for further feature selection.

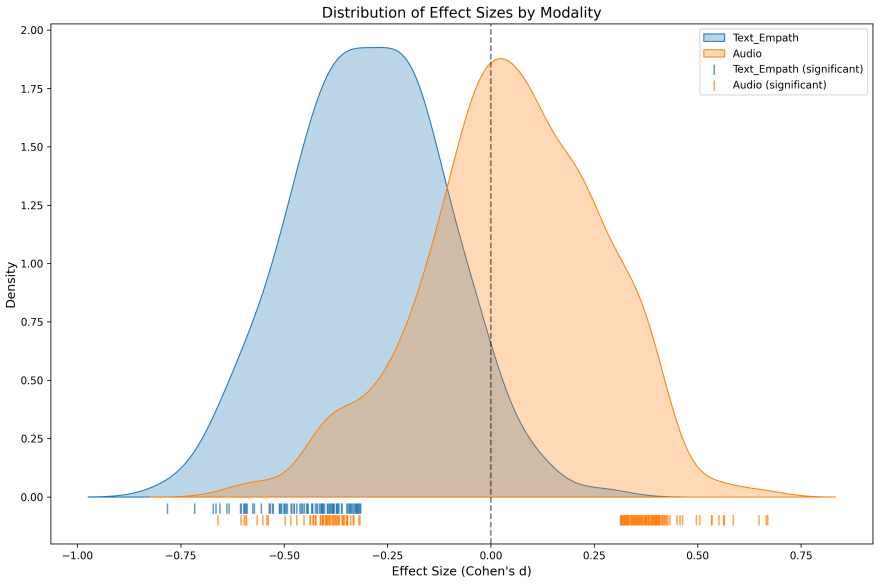

Figure 2. Effect Size Distribution of the Statistically Significant Features

After effect size filtering, only 45 features remain. Subsequently, a correlation matrix was generated to assess multicollinearity among these features. Highly correlated features were removed to reduce redundancy and improve model robustness. The correlation matrix before applying this correlation threshold can be found in figure 3.

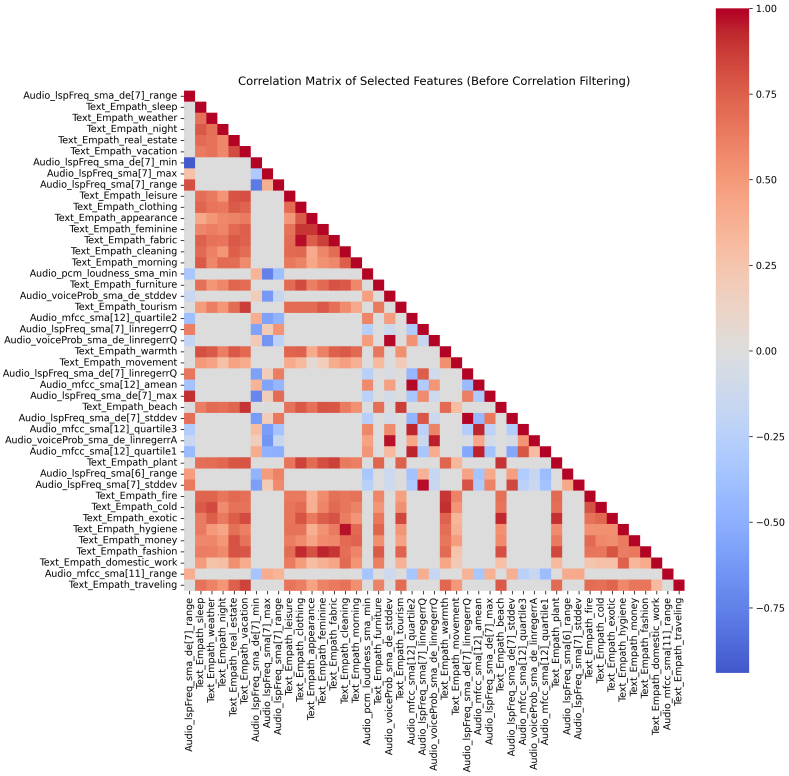

Figure 3. Correlation of the Selected Features Before Correlation Filtering

The correlation coefficients range from dark blue (-0.75) to dark red (0.75). The correlation of the features appears to be grouped into correlated blocks, with most blocks representing positively correlated features. A common correlation value is a correlation of approximately 0.5, particularly among features belonging to the same modality. The Empath categories such as *real estate* and *vacation* exhibit a correlation of approximately 0.75, suggesting that these semantic features frequently co-occur. Similarly, within the audio features, those belonging to the same category(i.e., *voiceProb*, *lspFreq*) tend to have a higher correlation. However, there are examples where this pattern does not hold. For example, the features *PCM loudness* and *voiceProb* show a correlation of approximately 0.5. This may be due to them reflecting the same aspect of speech, namely vocal effort. The *voiceProb* feature measures the probability of a voice presence in a certain audio segment, while *PCM loudness* quantifies the loudness or vocal intensity of the voice. These features are therefore intertwined and cannot occur separately. In contrast, the blocks with a near-zero correlation tend to contain a mix of textual and audio features, suggesting that these capture distinct, non-overlapping aspects of the data. In addition to examining these statistical interdependencies, potential hierarchical relationships between features were also assessed(e.g., features derived from one another, such as BMI from height and weight). No such hierarchies were found, so the assessment of interdependencies was based solely on correlation analysis.

The correlation matrix after applying correlation thresholding is shown in figure 4. The divide into multiple correlation blocks is more pronounced now. All remaining features are either uncorrelated or exhibit a moderate correlation(positive or negative) with each other. This pattern is particularly visible in the top left and bottom right corners of the matrix, which primarily contains features within the same modality, suggesting that they capture similar aspects of speech. This indicates that correlation thresholding has successfully reduced the redundancy within the feature set.

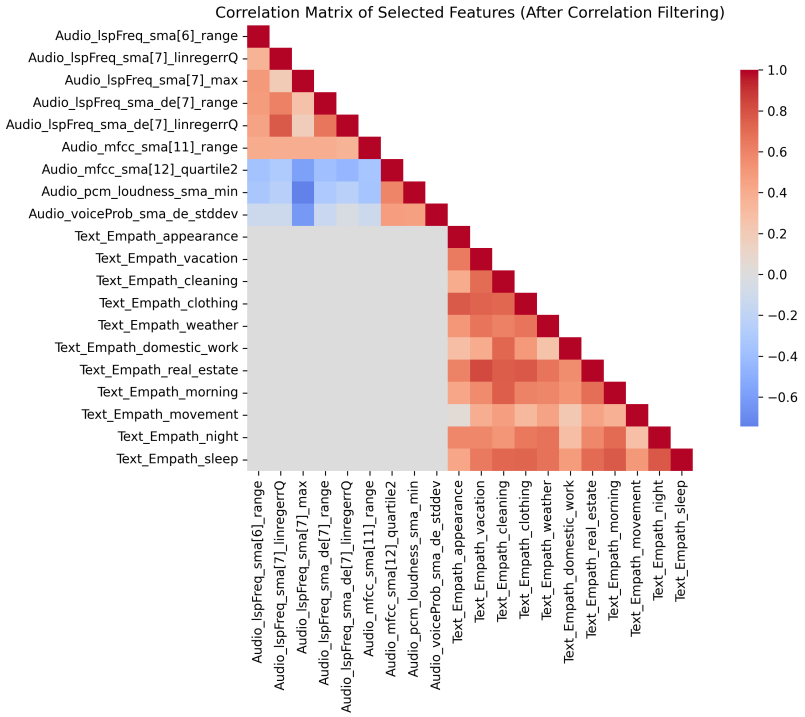

Figure 4. Correlation of the Selected Features After Correlation Filtering

4.2 Performance Metrics

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Random Forest	**0.6859**± 0.026	**0.6702**± 0.022	0.7326± 0.046	**0.6994**± 0.029	**0.7567**± 0.029
SVM	0.6778± 0.020	0.6612± 0.023	0.7319± 0.023	0.6944± 0.017	0.7311± 0.027
XGBoost	0.6741± 0.027	0.6582± 0.022	0.7237± 0.048	0.6889± 0.030	0.7551± 0.026
Linear Regression(Baseline)	0.6753± 0.008	0.6561± 0.005	**0.7394**± 0.013	0.6933± 0.009	0.7442± 0.009

Table 8. Performance Metrics of Classification Models

Table 8 shows the performance metrics of all the classification models, where for each metric, the best performance is boldfaced. The Random Forest model is the best performing model across all metrics, except for the recall metric. The differences between the best performing models for each metric compared to to the other models are marginal. All four models yield similar performance levels, with only slight variations in precision, recall, and F1 score. The ROC AUC scores of each model demonstrate that the models perform significantly better than random classification. The Random Forest model, the linear regression model and the XGBoost model have their highest score in the AUC metric. Furthermore, each model achieved a recall score of over 0.70, suggesting a strong ability to recognize the linguistic patterns associated with AD. To further illustrate the classification performance, the ROC curves for all models are presented in figures 5, 6,7, 8, showing the trade-off between true positive rate and false positive rate across different thresholds. The dashed line represents the performance of a random classifier.

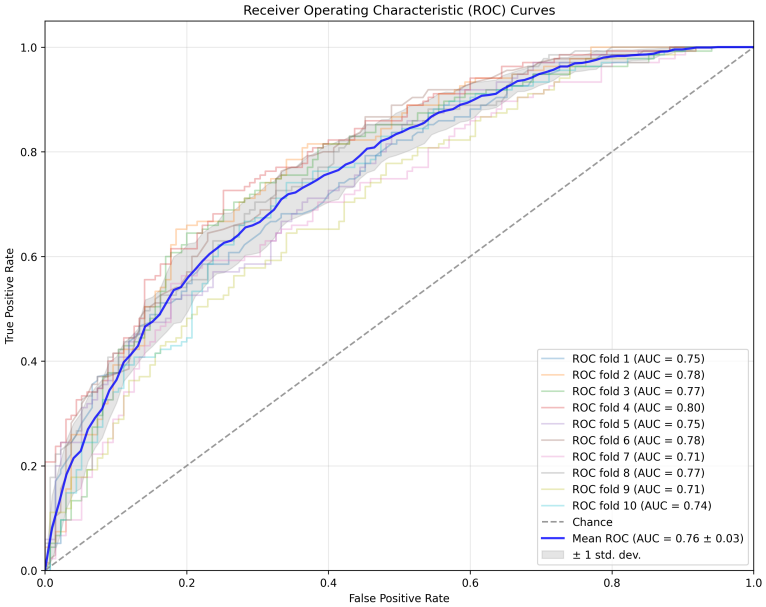

Figure 5. ROC Curve of the Random Forest Model

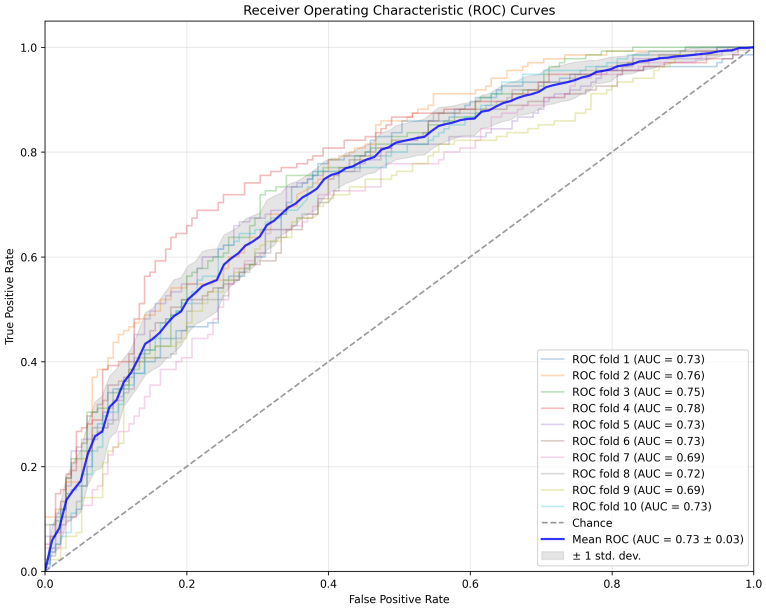

Figure 6. ROC Curve of the Support Vector Machine Model

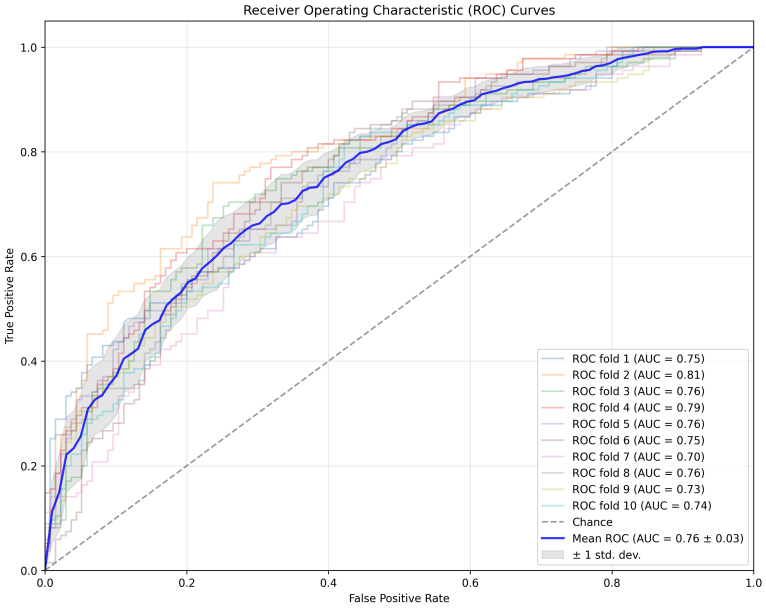

Figure 7. ROC Curve of the XGBoost Model

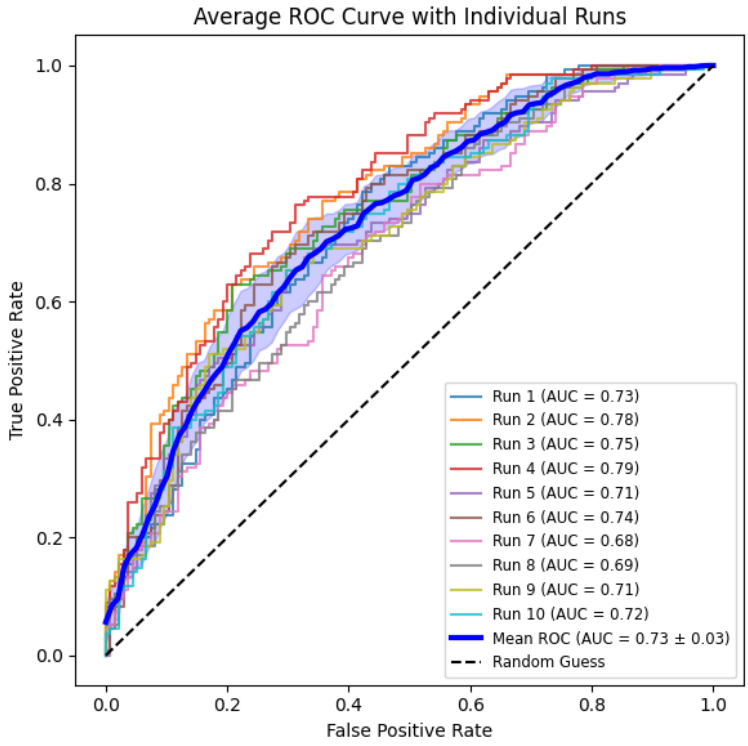

Figure 8. ROC Curve of the Linear Regression Model

The Random Forest model has a mean AUC score of 0.76(rounded to 2 decimal places). This suggests a good ability of distinguishing between the AD class and the control group. The mean AUC score has a standard deviation of 0.03, suggesting that the performance across the 10 runs remained similar. The consistent performance across folds suggests that the model is robust and not overfitting to any particular subset of the data. Similarly, the SVM and linear regression model had a mean AUC score of 0.73(SD = 0.03), and the XGBoost model reaches an AUC score of 0.76 (SD = 0.03). In all four models, the ROC curves for individual folds are closely clustered around the mean ROC curve, suggesting that the model's performance is consistent across different subsets of the data. These results demonstrate that all models generalize well and are not overfitting. The AUC values above 0.70 for all models indicate a fair to good

level of separability between AD and control classes. This consistency supports the overall stability of the classification framework and suggests that the models are not only able to capture relevant patterns associated with Alzheimer's Disease, but do so in a repeatable and dependable manner across different subsets of data.

4.3 Confusion Matrix Analysis

To gain a more detailed understanding of the predictions of the models, confusion matrices were generated for each model. These matrices provide insight into how well the models distinguish between the AD and control groups by analyzing the true positives, false positives, true negatives, and false negatives. This enables the evaluation of not only the overall performance but also class-specific strengths and weaknesses. The confusion matrix of the Random Forest model is shown in figure 9. The true labels are depicted on the y-axis and the predicted labels on the x-axis. The diagonal cells represent correct predictions and the off-diagonal cells the misclassifications.
The Random Forest model correctly predicted 86 cases of Alzheimer's Disease and 99 control cases. However, it incorrectly classified 49 AD cases as control and 36 control cases as AD. The Random Forest model is more effective at correctly classifying AD cases than control cases. This is further supported by the sensitivity and specificity. The sensitivity (**True Positive Rate**) for the AD class is Sensitivity $= \frac{86}{86+49} \approx 0.64$, indicating that the model correctly identifies about 64% of AD cases. The specificity (**True Negative Rate**) for the control group is $\frac{99}{(99+36)} \approx 0.73$, indicating that approximately 73% of the control cases are correctly identified. The False Negative Rate(**FNR**) is $\frac{49}{(86+49)} \approx 0.36$ and the False Positive Rate(**FPR**) is $\frac{36}{(99+36)} \approx 0.27$.

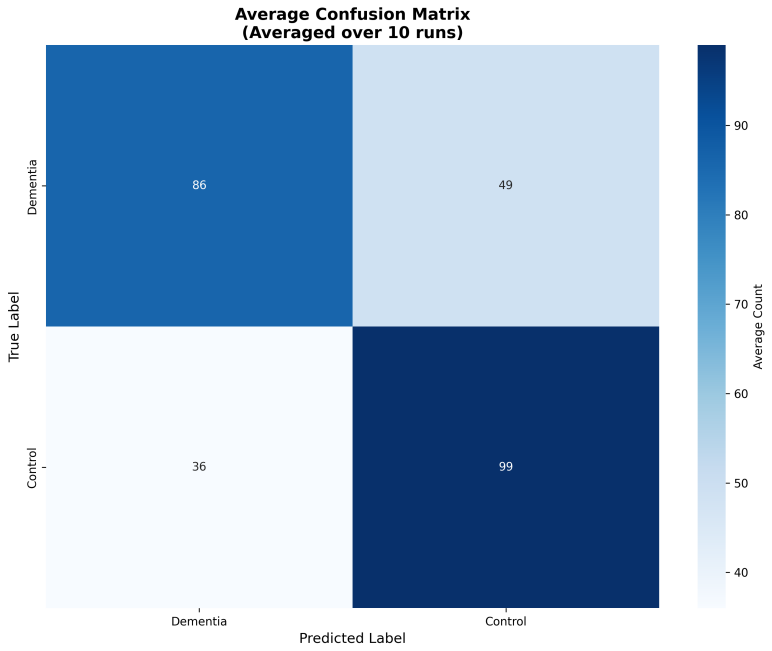

Figure 9. Confusion Matrix of the Random Forest Model

The SVM model correctly predicted 84 AD cases and 99 control cases, as illustrated in figure 10. It misclassified 51 AD cases as control and 36 control cases as AD. The sensitivity is $\frac{84}{84+51} = 0.62$ and the specificity is $\frac{99}{99+36} = 0.73$, indicating a similar performance to the Random Forest model. It correctly identified 62% of AD cases and 73% of the control cases. Overall, the model seems to perform well. The FPR is $\frac{36}{36+99} = 0.27$ and the FNR $\frac{51}{84+51} = 0.38$. Similar to the Random Forest model, the SVM model performs slightly better on the control cases than on AD cases.

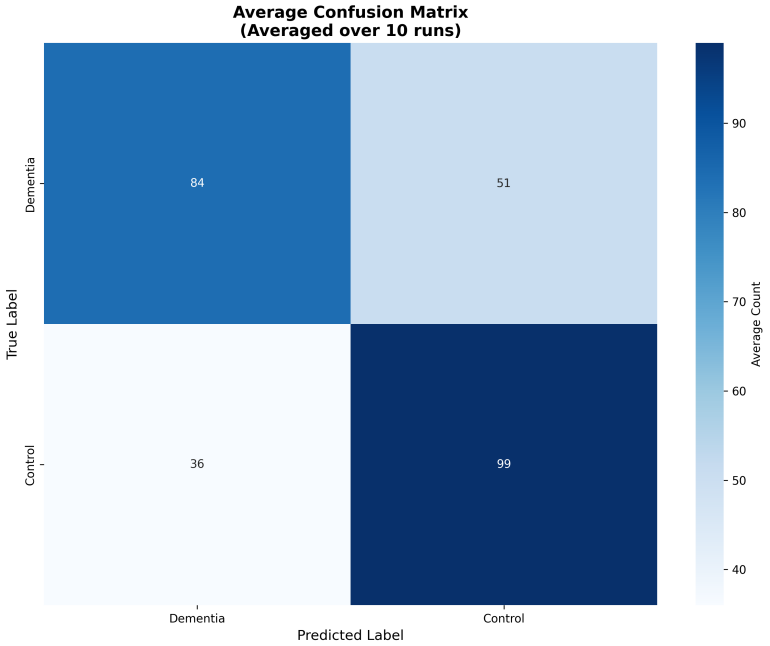

Figure 10. Confusion Matrix of the Support Vector Machine Model

The confusion matrix of the XGBoost model in figure 11 reveals that it correctly classified 84 AD cases and 98 control cases. The sensitivity is $\frac{84}{(84+51)} \approx 0.62$ and the specificity $\frac{98}{(98+37)} \approx 0.73$. The FPR is $\frac{37}{(98+37)} \approx 0.27$ and the FNR is $\frac{51}{(84+51)} \approx 0.38$. Similarly to the previous models, the XGBoost model appears to predict the control cases better than the AD cases.

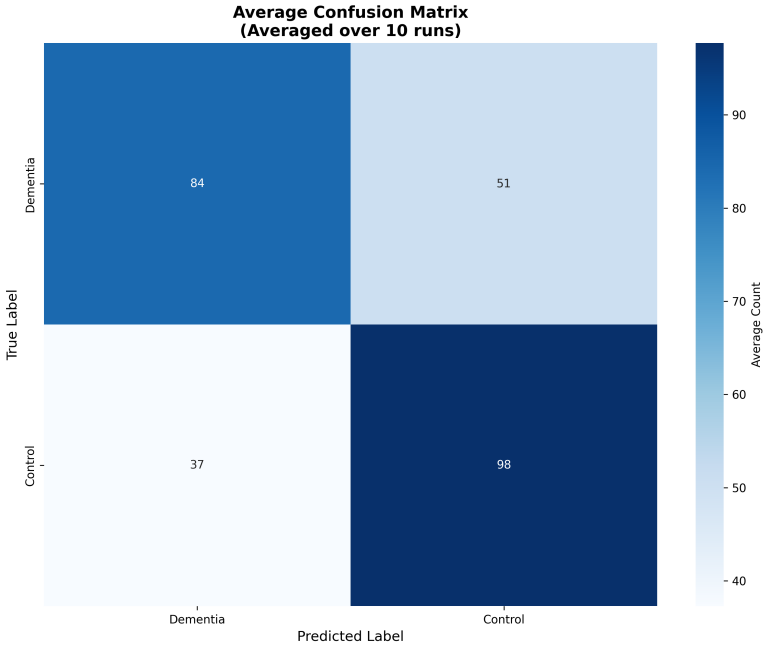

Figure 11. Confusion Matrix of the XGBoost Model

Finally, the confusion matrix of the linear regression model in figure 12 shows it correctly classifying 81 AD cases and 99 control cases. It misclassified 54 AD cases as control and 36 control cases as AD. The sensitivity is $\frac{81}{(81+54)} = 0.60$ and the specificity $\frac{99}{(99+36)} \approx 0.73$. The FPR is $\frac{36}{(99+36)} \approx 0.27$ and the FNR is $\frac{54}{(81+54)} \approx 0.40$. The linear regression model, in line with the other three models, appears to predict the control cases better than the AD cases.

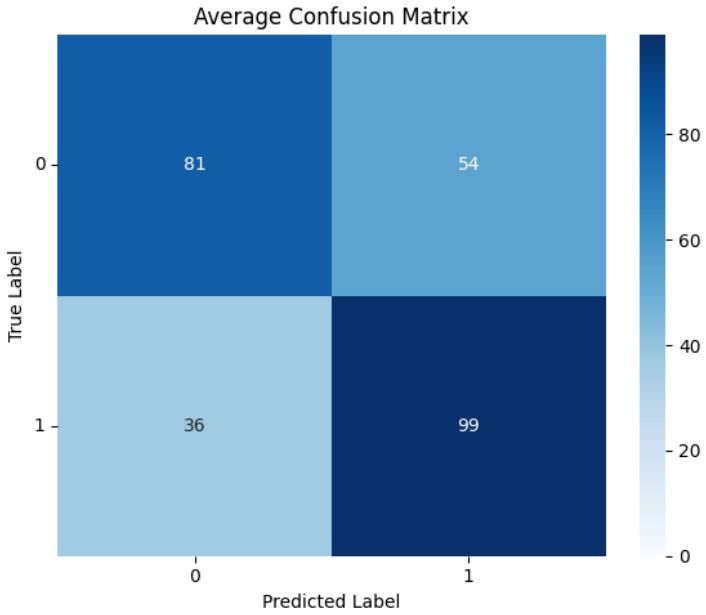

Figure 12. Confusion Matrix of the Linear Regression Model

4.4 Frequency Analysis of Selected Hyperparameter Combinations

This subsection discusses the hyperparameter values that were selected during the 10 runs for all models, excluding the baseline model. The first model to be examined is the Random Forest model. The frequencies of the selected hyperparameter values for this model are shown in figure 13.

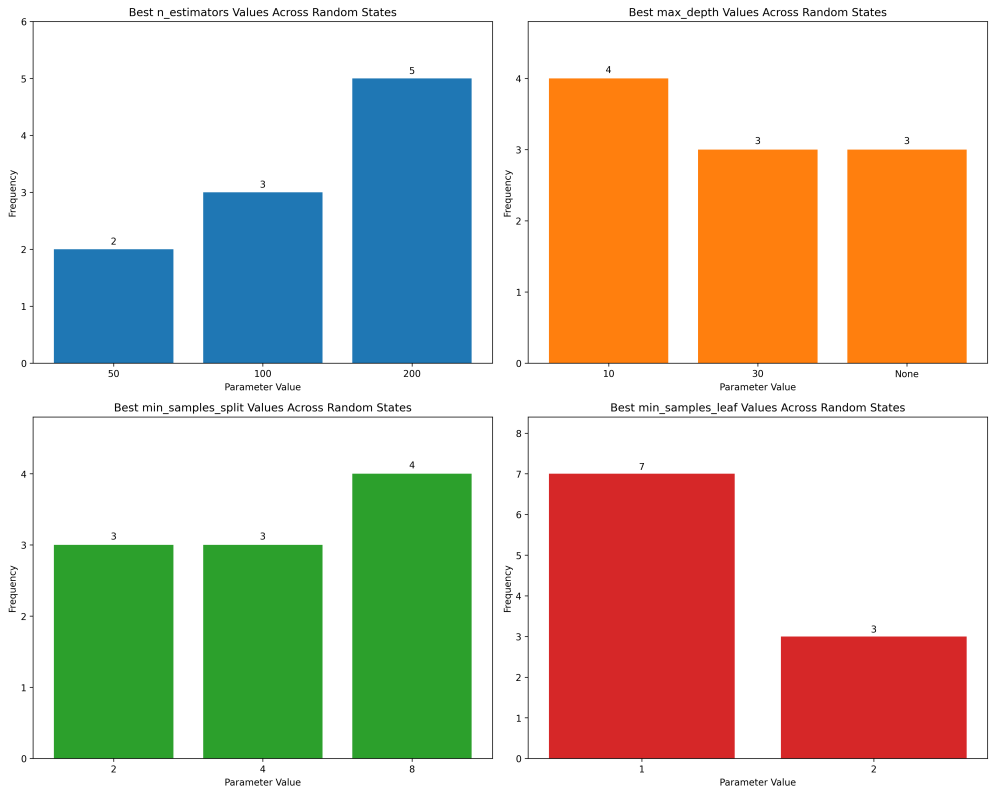

Figure 13. Hyperparameter Frequencies of the Random Forest Model

The tuned hyperparameters for the Random Forest model were: `n_estimators`, `max_depth`, `min_samples_split`, and `min_samples_leaf`. The hyperparameter `n_estimators` defines the

number of trees in the forest. When increased, it can improve the model's performance by reducing variance, but is computationally more expensive. `max_depth` sets the maximum depth of each decision tree. Limiting this can help prevent overfitting. The third hyperparameter, `min_samples_split`, refers to the minimum number of samples required to split an internal node. When this is decreased, the size of each decision tree increases. Similarly, `min_samples_leaf` sets the minimum number of samples each leaf node has. This can also make the tree larger when this value is small and small when the opposite occurs. The value that was most frequently selected for `n_estimators` was '200', followed by '100' and '50'. This suggests that complexity of the task warrants using more trees. For `max_depth`, the value '10' was selected most often in 4 out of 10 runs, followed by '30' and 'None'. Both selected in 3 out of 10 runs. The small difference in the selection frequency suggests that the model has no strong preference for any hyperparameter value. For `min_samples_split`, the value '8' was selected the most in 4 runs, while '4' and '2' were each selected in 3 runs. The model again shows no clear preference for any hyperparameter value. In contrast, for `min_samples_leaf`, the value '1' was chosen in 7 out of 10 runs and '2' in 3 runs, clearly favoring smaller leaf sizes and deeper trees.

For the SVM model, the tuned hyperparameters were `C`, `kernel`, `degree`, and `gamma`. The hyperparameters and the frequency at which the values were chosen, can be found in figure 14.

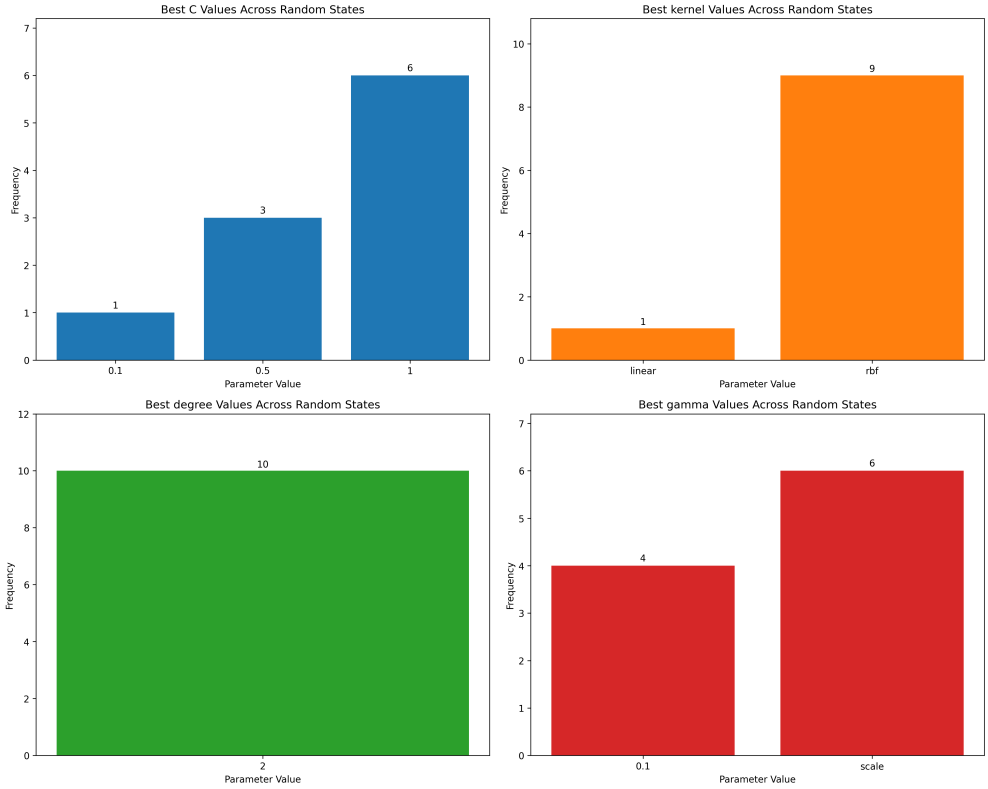

Figure 14. Hyperparameter Frequencies of the Support Vector Machine Model

`C` controls the trade-off between a low training error and a low testing error, `kernel` determines the function used to transform the feature space. The `degree` hyperparameter controls the degrees of the polynomial function and `gamma` controls the influence each training sample has on the decision boundary. A higher value for `gamma` means that it will only consider points nearby the decision boundary, potentially leading to overfitting. A lower value means that it will consider samples that are further apart, leading to a smoother decision boundary which may result in underfitting. The most frequently selected value for `C` was '1', followed by '0.5' and '0.1'. This may indicate that the model tried to balance the optimization of the training and test error. The 'rbf' kernel was selected in 9 out of 10 runs and the 'linear' kernel once. This indicates that the model has a strong preference for the 'rbf' kernel over the 'linear' kernel. For `degree`, the value '2' was chosen in all 10 runs. Regarding `gamma`, the

'scale' value was selected in 6 out of 10 runs, while the float value '0.1' was chosen in 4 runs. The 'scale' value for `gamma` adjusts the influence each training sample has based on the dataset. A value of '0.1' means that training samples have far-resulting influence, leading to a more smooth decision boundary.

The hyperparameters tuned for the XGBoost model were: `learning_rate`, `max_depth`, `gamma`, and `subsample`. These can also be found in figure 15.

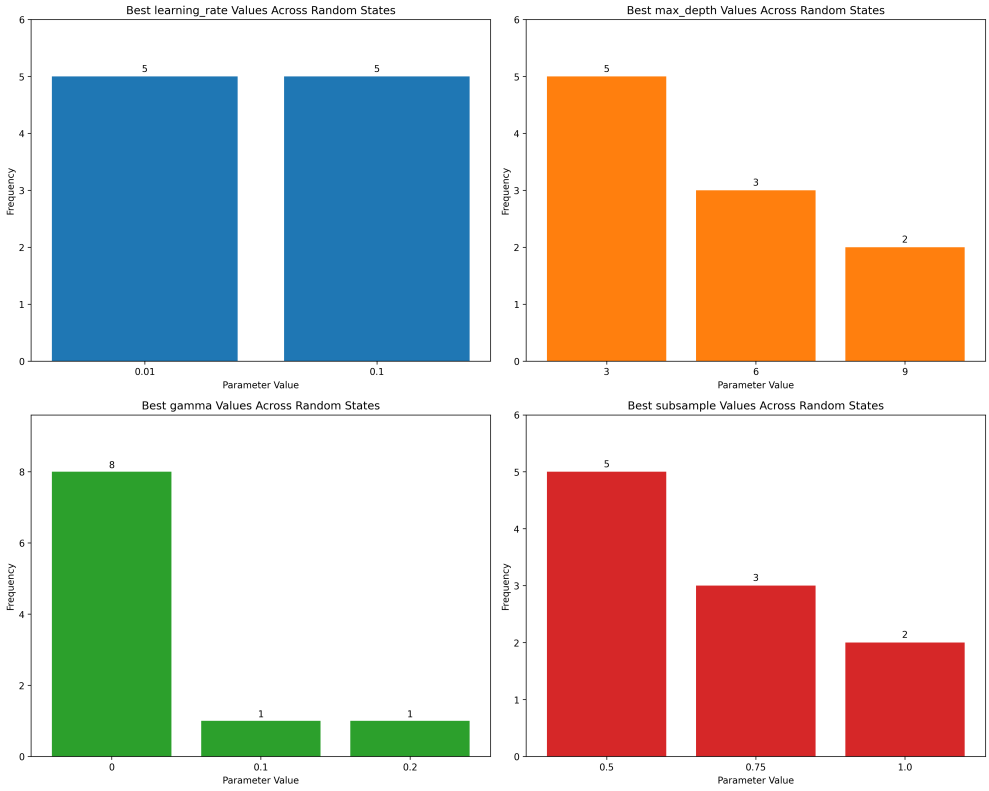

Figure 15. Hyperparameter Frequencies of the XGBoost Model

The `learning rate` controls the speed at which the model learns. A smaller learning rate value forces the model to learn more gradually and makes it more robust against overfitting, but requires more trees to converge. `max_depth` controls the depth of each individual tree. A higher value makes the tree deeper and more able to capture complex patterns, at an increased risk of overfitting. `gamma` sets the minimum reduction of the loss required for a node to split. A higher `gamma` value would lead to fewer splits and smaller trees, while a lower `gamma` value would do the opposite. `subsample` defines the proportion of the training data used to train each tree of the model. A lower value may introduce more variation which may decrease the risk of overfitting, but also has the risk of underfitting. A higher value for `subsample` may increase the risk of overfitting. The values '0.01' and '0.1' for the `learning_rate` were both selected for 5 runs, while the value of '0.3' was not selected for a single run, suggesting a preference for a lower learning rate and slower convergence. For `max_depth`, the value '3' was selected in 5 runs, followed by '6' in 3 runs and '9' in 2 runs. This may indicate that a `max_depth` of '3' offers the best balance between variance and bias. The `gamma` value of '0' was chosen 8 times, followed by '0.1' and '0.2' that were chosen ones each. This shows that the model may perform better when the tree has fewer inhibitors to grow. The last hyperparameter `subsample`, follows a similar distribution of `max_depth`. The value '0.5' was selected for 5 runs, '0.75' for 3 runs and '1' for 2 runs. This may imply that using half of the training data for each tree yields the best performance.

4.5 Feature Importance Analysis

This section covers the feature importance of the selected features across all four models. First, the feature importance for the Random Forest model will be discussed. The feature importance scores for this model are shown in figure 16.

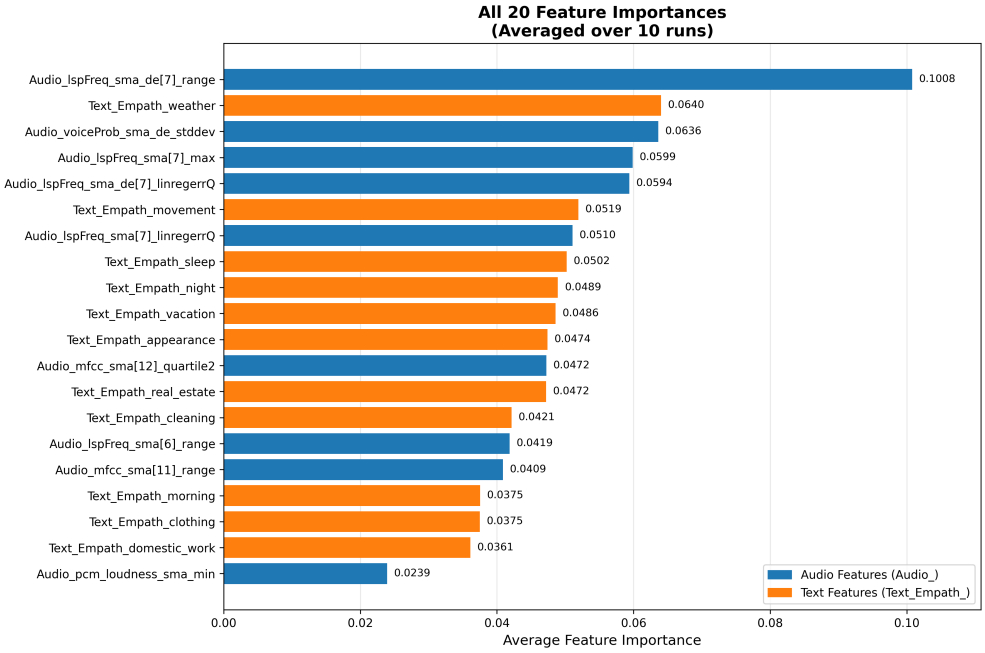

Figure 16. Feature Importance of the Selected Features of the Random Forest Model

The top five most important features are primarily audio features, with only one text feature included. There is a significant difference in the feature importance of the most important and second most important feature. Beyond this, the difference between the feature importance scores gets significantly smaller, tapering off towards the two lowest-ranked features. Overall, audio features appear to be more influential than the text features. The model seems to rely more heavily on audio features, particularly those related to frequency and voice probability, to make predictions. The text features still contribute meaningfully, particularly the categories related to environmental and movement contexts, though are less influential. This may suggest a more balanced approach, where both audio and text features add value, but audio features play a more critical role in the model's performance. While audio features were more influential, the model's use of specific text categories still reveals meaningful semantic patterns warranting further analysis. To better understand how specific Empath categories were activated by the words of the participants, a script was run that computed the cosine similarity between each individual word of the participant and each Empath category label. The ten words that had the largest cosine similarity for each category were extracted to guide further interpretation. For instance, the *weather* category may be linked to the water overflowing the sink because the word 'flooding' was used to describe it, which falls under the *weather* category. Similarly, the *movement* category most likely refers to the boy standing and almost falling off the stool, through words such as 'falling' and 'tumbling'. The *sleep* and *night* category might have been triggered by a phrase as: 'The mother is daydreaming', since the word 'dreaming' falls under the *sleep* category and 'daydreaming' under the *night* category. The *vacation* category could be associated with phrases such as 'like a summer afternoon', since "summer" is directly tied to vacation-related contexts. Additionally, the word 'lalaland' used to describe the mother was not paying attention might contribute to the *vacation* category due to their metaphorical connection to being mentally 'away' or 'not present'. The second model to consider is the SVM model, for which the feature importance scores are displayed in figure 17.

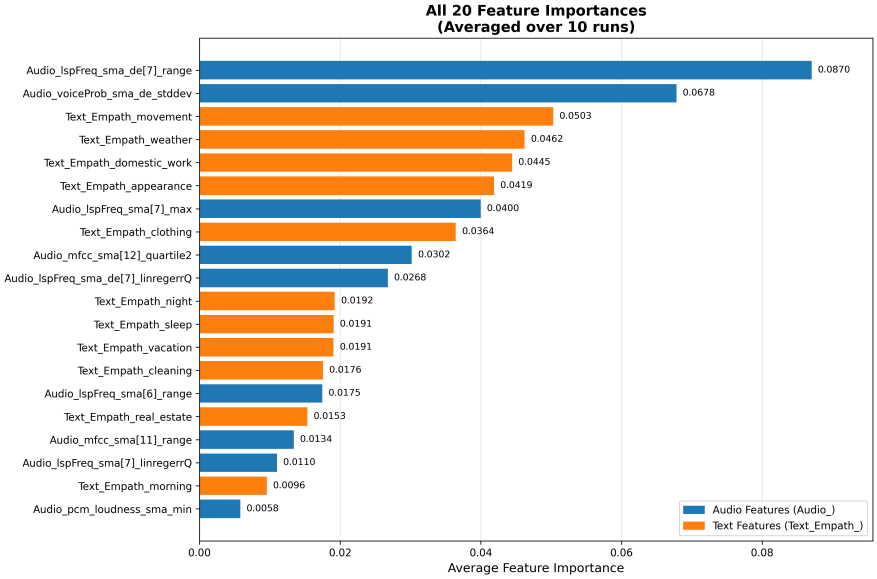

Figure 17. Feature Importance of the Selected Features of the Support Vector Machine Model

The SVM model does not have native feature importance scores, so these were derived using permutation importance. Permutation importance is calculated by shuffling the value of a feature across all samples and measuring the drop in performance. A bigger drop in performance indicates a higher importance. The top five most important features are primarily text features, although less influential than the audio features, that are occupying the first and second place. The audio features generally have higher feature importance scores than text features. Similarly to the previous model, this relies slightly more on audio features than text features to make a prediction. The most important audio features seem to relate to the voice frequency and voice probability. Important text features related to movement, the *weather* and *domestic work* categories.

The feature importance scores for the XGBoost model are shown in figure 18.

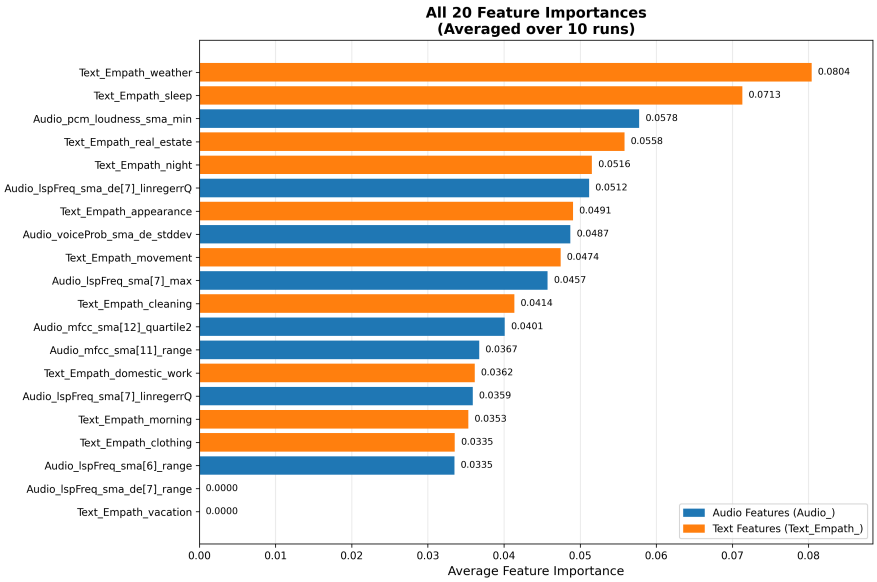

Figure 18. Feature Importance of the Selected Features of the XGBoost Model

In contrast to the previous two models, the top five most important features are dominated by text

features, with the weather category being the most important. Text features generally appear to be more important than audio features. Textual features such as the *weather*, *sleep*, and *real estate* categories seem to be the most important. For the audio features, those related to voice frequency, voice probability and loudness are more prominent. This is also the first model where some features have an importance score of zero, indicating they do not contribute to the model's predictions.
Finally, the feature importance scores of the baseline model are displayed in figure 19.

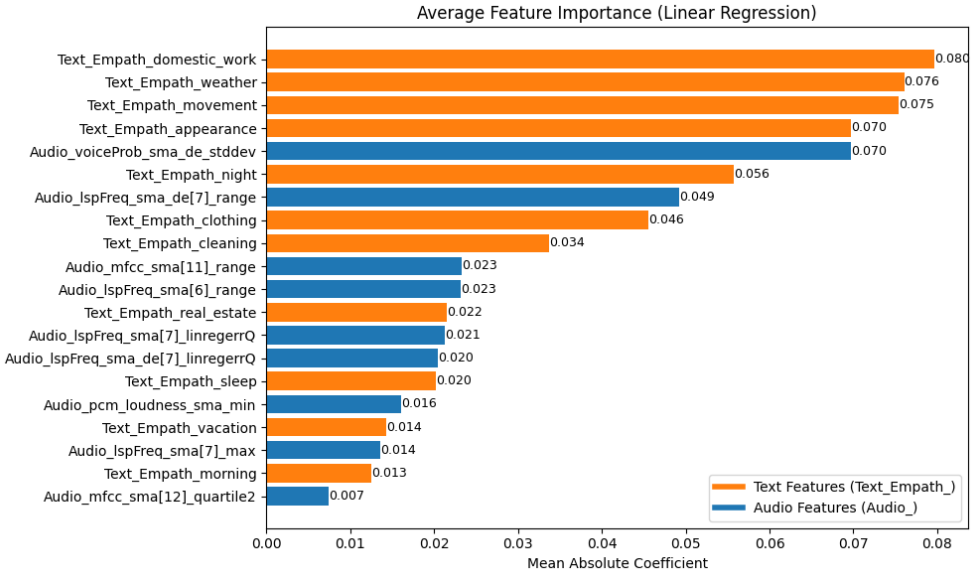

Figure 19. Feature Importance of the Selected Features of the Linear Regression Model

The five most important features consist mostly of text features, comparable to the SVM and XGBoost model. The most important feature is the *domestic work* category, followed by the *weather* category, the *movement* category, the *appearance* category and the `voiceProb` audio feature. In general, the text features seem more influential than the audio features. Feature importance scores vary little within defined feature index ranges (1–5, 6–8, 9, 10-15, 16-19, 20), but differ more between them. The five most important features for each model are summarized in table 9, in the order of most important to fifth most important.

Linear Regression(Baseline)	**Random Forest**	**Support Vector Machine**	**XGBoost**
Domestic Work	range of lspFreq	range of lspFreq	Weather
Weather	Weather	voiceProb	Sleep
Movement	voiceProb	Movement	PCM Loudness
Appearance	max value of lspFreq	Weather	Real Estate
voiceProb	linregerrQ lspFreq	Domestic Work	Night

Table 9. The Five Most Important Features for each Model

5 DISCUSSION

This study began by using statistical analysis to identify relevant features, after which three machine learning models and a linear regression model(baseline) were trained and evaluated. The results provided insight into how audio and text-based features contributed to model performance. Moreover, they highlighted the different ways in which features were utilized, resulting in varying feature importance scores and overall model performance. The results demonstrated consistent patterns across performance metrics, such as three out of four models achieving their highest scores in the ROC AUC metric. The confusion matrices also revealed a recurring pattern: all four models were more accurate in predicting control cases than cases of Alzheimer's Disease. Furthermore, hyperparameter tuning highlighted the impact different values for the hyperparameters had on the performance metrics, and in some cases, demonstrated clear preferences for specific hyperparameter settings. Collectively, these findings not only validate the effectiveness of the selected features, but also uncovered the different manners in which each model leveraged audio and text data to make predictions. This offers valuable guidance for future model development and feature engineering in similar contexts.

5.1 Effectiveness of Statistically Selected Features in Model Performance

This section discusses the results of the feature selection based on statistical analysis and its impact on the performance metrics. The statistical analysis began by retaining only the features that were statistically significant. Of the remaining features, those with an effect size below 0.5 were removed. To reduce multicollinearity, features with a correlation above 0.8 underwent further filtering: the feature with the higher effect size was retained. If the effect sizes were equal, the feature with the higher variance was selected. In the event of a tie, the feature with the earlier alphabetical name was selected.

The effectiveness of this feature selection method is demonstrated by the consistent performance of all four models across the performance metrics. Performance differences between models were marginal, and all four models achieved moderately high ROC AUC scores. This suggests that the features were both informative and relevant for the classification task. In medical contexts such as disease recognition, recall is crucial for identifying as many of the true positive cases as possible [50]. All four models achieved recall scores of between 0.72 and 0.74, which aligns with results reported in prior studies. The recall in the literature often ranges from 0.68 to 0.84[46][66][34]. While this study employed machine learning models to classify AD, it is important to note that the previously discussed studies used different methodologies. For instance, Miller et al. (2014) used MRI images as input for AD classification, while Twait et al. (2023) used a combination of MRI images and biometric assessments. Prior studies have often shown that text-based feature models consistently outperform those based on audio in Alzheimer's Disease classification tasks [57, 37]. Based on this, it was expected that text features would play a more significant role in model performance. This expectation was not met. This may be due to the utilization of different audio features. Lin and Washington (2024) extracted features using Wav2Vec. Wav2Vec, in contrast to OpenSMILE, does not use preprocessed features but directly operates on raw audio[60]. It works by trying to predict missing parts of the audio based on the surrounding sounds[60]. The emobase feature set from OpenSMILE is tailored towards recognizing emotions. The feature set has features for measuring certain auditory indicators of Alzheimer's Disease such as a flat tone in their voice(`lspFreq`) and measuring the pause between words and/or sentences(`voiceProb`, `PCM loudness`), but not every auditory signal of AD. This may make Wav2Vec more suitable for AD detection than the emobase feature set of OpenSMILE.

Another factor potentially influencing the performance of the model could be the lack of data augmentation in this study. Lin and Washington(2024) applied data augmentation, which may have improved the robustness and diversity of their model. The absence of data augmentation in this study could have limited the performance. Another unexpected yet informative result was the near-zero correlation between most text and audio features in the correlation matrices. This implies that the different modalities capture distinct aspects of speech, which supports the case for multimodal modeling. Finally, the performance of the three machine learning models, both compared to each other and to the baseline model, was similar. It was anticipated that one model would outperform the others and that the machine learning models would significantly outperform the baseline model. This could be attributed to the limited size of the dataset. The dataset used in this paper had 674 samples. A study found that the XGBoost model reaches its AUC peak at approximately 9960 samples and 3404 for the Random Forest model[62]. This study may not have had the

sample size needed to measure significant differences in the performance of the three models to each other and the baseline model. Furthermore, the sample size may have caused the more complex models to overfit.

In conclusion, the method of statistical feature selection demonstrated its effectiveness in identifying relevant features, as evidenced by the consistent performance of all four models across key metrics. The achieved recall scores (0.72–0.74) align with prior studies, reinforcing the practical utility of the selected features for AD classification. It was expected that text-based features would dominate performance, which was not the case. The audio features contributed more than expected, potentially influenced by the use of OpenSMILE's emobase feature set, which emphasizes emotion-related auditory indicators, versus the Wav2Vec framework employed in other studies. Additionally, the absence of data augmentation in this study may have limited feature robustness compared to other studies. The near-zero correlation between text and audio modalities further underscores their complementary nature, advocating for multimodal approaches. The similar performance among the three machine learning models as well as in comparison to the baseline model, likely stems from the dataset's limited size (674 samples), which may not have provided sufficient statistical power to fully distinguish model capabilities. These findings highlight the interaction between the feature selection method, modality-specific extraction techniques, and dataset constraints in shaping model outcomes.

5.2 Interpreting Class-wise Performance Through Confusion Matrix Analysis

To gain deeper insight into the model's behavior beyond aggregate performance metrics, this section will analyze the results of the confusion matrix analysis. This enables the examination of class-wise predictions. While performance metrics such as accuracy and ROC AUC provide a broad understanding of the model's performance, the confusion matrix offers critical insight into how well each class was distinguished. This is particularly important for identifying AD cases. In the context of AD detection, understanding the number of misclassified AD cases can open new paths for further research, given that an early and correct diagnosis of AD is beneficial. The Random Forest model correctly predicted 64% of the AD cases and 73% of the control cases. Both the SVM and XGBoost models correctly predicted 62% of the cases of Alzheimer's Disease and 73% of the control cases. Lastly, the baseline model predicted 60% of the AD cases and 73% of the control cases correctly. All three machine learning models and the baseline model consistently performed better on control cases than on AD cases. The consistency of this pattern across all models suggests that the issue likely stems from the input features and class characteristics rather than specific inherent limitations of any of the models.
Hernández-Domínguez et al. (2018) achieved a sensitivity of 87%. They took 25 descriptions of the Cookie Theft picture from healthy older adults and used those to create a baseline or "ideal" version of what a good, informative answer looks like[29]. Then, they evaluated 517 samples, 257 AD samples, 217 control samples, and 43 samples of individuals with mild cognitive impairment[29]. Their linguistic features included the total word count, number of unique words, and word frequency counts[29]. For the audio features, they only used the first 12 to 13 coefficients of the Mel-Frequency Cepstral Coefficients (MFCCs), based on the assumption that the first 12 to 13 capture the most important patterns. MFCCs measure how the sound energy is distributed across frequencies. They checked how similar and informative each person's description was compared to the model answer[29]. More focus was on information coverage, where the researchers checked if the participant omitted important details, focused on irrelevant aspects, and clearly referred to people and objects(e.g., "the boy" vs. "he")[29]. This study achieved a lower sensitivity, which may be explained by their use of features specifically tailored towards detection of AD. Another study achieved a sensitivity of 76% with the Random Forest model[11]. Their text features focused on measures of speech quantity (e.g., total words, total number of nouns) and fluency (e.g., filler words(um, uh, hmm))[11]. They also assessed the use of specific word types (e.g., definite/indefinite articles, verbs)[11]. These features are a reflection on how varied, informative and fluent the language of the participants was. These features were specifically tailored towards detecting AD, whereas the sentence transformer employed in this study is intended for general use. Additionally, the absence of audio features may explain the difference.

The number of AD and control samples were equal in the present study, ruling out class imbalance as an explanation for the disparity in performance. One possible explanation lies in the variability of speech among AD patients. A study that analysed 1000 voice samples found that those with AD had moderate variety in mean speech segment duration and high variety in total duration of speech, mean

pause duration, and speech rate[58]. This variability suggests that AD affects individuals vocal features differently, leading to a broader range of speech behaviors and patterns. As a result, machine learning models may struggle to generalize, particularly when trained on relatively small datasets. Unlike more uniform speech patterns typically observed in control subjects, the acoustic diversity of patients with Alzheimer's Disease may prevent models from learning consistent, discriminative features. This could make it harder to capture underlying patterns of speech in AD patients. Another potential explanation may lie in how the emobase feature set works. The features in emobase are derived by using statistical functions, such as the mean and the standard deviation, on 26 Low Level Descriptors[16]. This results in a summary of the audio signal rather than a detailed account of how the signal changes over time.
While this may suffice for emotion recognition tasks, it could have made the model miss subtle details that could have been important for distinguishing the AD group from the control group. This disparity in performance also raises concerns about potential bias in the models. The tendency to perform better on control cases than on AD cases may indicate that the models are more attuned to structured or typical speech patterns. This has severe implications from an ethical perspective. A false negative, where an AD case is misclassified as cognitively healthy, can delay diagnosis and further intervention. These are not merely statistical errors, but can have severe consequences on affected individuals.

In conclusion, the confusion matrix analysis highlights key areas where the models fall short, particularly in detecting language patterns associated with AD. The models in this study underperformed when compared to those in the literature. This underscores the importance of specifically tailoring the features towards detecting AD. It also reinforces the need for larger, more diverse datasets and careful attention to ethical considerations in clinical applications. To further unpack where the models went wrong, an error analysis was conducted. Across all four models, sentences such as *'the mother is dryin(g) a plate.'*, *'the mother's oblivious to all.'*, *'the woman is, the the mother is washing drying the dishes.'* and *'the sink, the the faucet's on'* contributed to a significant number of misclassifications. For the first two sentences, it appears that the models classified too many short, telegraphic- like sentences as AD. While it is correct that patients with AD tend to use shorter sentences more frequently than healthy individuals, this feature is not unique to AD. In fact, in the Cookie Theft picture description task, even healthy controls often produce short, simple sentences, since the task encourages brief descriptive utterances rather than complex narratives. The models therefore seem to have over-associated short sentences with AD. This may have been due to the limited number of samples. For the last two sentences, the transcripts show that participants sometimes restart their description. For example, one individual began with *'the woman is'*, but then corrected it to *'the mother is'*. Similarly, in the final example, *'the sink'* was corrected to *'the faucet'*. Individuals with Alzheimer's Disease are more likely to restart and self-correct, which can be noted as a disfluency, which is a marker for AD. However, in these cases, the models may have over-weighted these disfluencies as evidence of AD.

5.3 Implications of Observed Hyperparameter Frequencies

This section discusses the implications of hyperparameter frequencies observed across the three machine learning models. Understanding why and which hyperparameter values were frequently selected can help prioritize tuning efforts in future work. For example, if the learning rate consistently requires fine-tuning across models, it likely plays a crucial role in model performance. The tuned hyperparameters for the Random Forest model were: `n_estimators`, `max_depth`, `min_samples_split`, and `min_samples_leaf`. For the SVM model, they included `C`, `kernel`, `degree`, and `gamma`.
The XGBoost model used `learning_rate`, `max_depth`, `gamma`, and `subsample`.

The first hyperparameter to be discussed is one that appears in both the Random Forest and XGBoost models. A comparison of `max_depth` across models revealed that non-default values were selected as the best option in 70% of cases for both Random Forest and XGBoost. While the Random Forest model showed no strong preference for any of the non-default values for the `max_depth` hyperparameter, XGBoost clearly preferred shallower trees (value 3) compared to the default (value 6). This result suggests that `max_depth` is an important hyperparameter to tune for both models. For this dataset, simpler trees worked better for XGBoost, while Random Forest was more flexible regarding tree depth. Additional hyperparameters controlling the complexity of the trees include `min_samples_split`, and `min_samples_leaf` for the Random Forest model and `gamma` for the XGBoost model. The default value '0' for `gamma` was chosen 8 out of 10 runs. For the XGBoost model, the default value of `gamma`

increases the size of each individual tree. However, the tree depth was already constrained in this study, with the max_depth hyperparameter set to a value of 3 in 5 out of 10 runs. The alternative values '6' and '9' also represent shallow trees. Given these constraints, additional restrictions on the tree size via gamma were redundant. For the Random Forest model, the default min_samples_split=2 was selected in 3 out of 10 runs, meaning that 7 runs do not pick the default value, whereas min_samples_leaf=1 (default) was selected in 7. For two out of three hyperparameters controlling tree complexity, the model strongly preferred the default value. This suggests that shallow splits could have often led to many small, unstable nodes that resulted in overfitting, while small leaf sizes did not.

For SVM, the most frequently chosen value for C was 1 (default), selected in 6 out of 10 runs. Lower values such as 0.5 and 0.1 were chosen in 3 and 1 run, respectively. The SVM shows a strong preference for the default value, which leads to moderate regularization. Lower values for C allow for more classification errors which leads to a higher margin, while higher values prioritizes a higher training accuracy leading to a narrower margin. The remaining hyperparameters to be discussed are learning_rate, subsample, and the kernel hyperparameter.
The learning_rate values '0.01' and '0.1' for the XGBoost model were both chosen 5 times, while the default value '0.3' was never selected. The dataset may have had more complex patterns or noise that required more gradual and more controlled learning. A higher value such as '0.3' may lead the model to fail to converge. Similarly, for subsample, values of 0.5, 0.75, and 1 were selected 5, 3, and 2 times, respectively, with 0.5 being most frequently chosen.
The default value was chosen the fewest number of times. That may be due to the values 1 and 0.75 leading to overfitting. Subsampling for the value 0.5 likely acted as a stochastic regularizer leading to a more diverse training set. This, in conjunction with shallower trees, may have contributed to a more favorable bias-variance tradeoff. The last hyperparameter to be discussed is the kernel hyperparameter of the SVM model. The 'linear' and 'rbf' values were chosen 1 and 9 times respectively. This suggests the data exhibited non-linear patterns, which the RBF kernel could better capture. The linear kernel was chosen once. This may be due to a specific train-test split where classes were able to be separated linearly. Since the poly kernel was never selected, the degree hyperparameter became irrelevant in practice.

For the Random Forest model, the min_samples_split and max_depth hyperparameter fluctuate the most from their default values, suggesting they are highly sensitive and require tuning to avoid overfitting. In the SVM model, degree varied the most from the default value due to the 'poly' kernel not being chosen for a single run, which nullifies its influence. The most sensitive hyperparameter for the XGBoost model is subsample. The default value was chosen in 2 out of 10 runs. As previously mentioned, this may have been due to trying to prevent the model from overfitting. The overall preferred values for the Random Forest model suggest an emphasis on controlling the tree complexity in addition to limiting the tree depth(max_depth=10). This is in combination with trying to prevent overfitting by making the requirements to make splits higher. This means that the dataset likely has some complexity that benefits from many trees (n_estimators=200), but also comes with a risk of overfitting that requires constraints on the size of each tree, hence the higher values for min_samples_split and min_samples_leaf.
The SVM model leaned towards the default value '1' for C. and strongly preferred the non-linear rbf kernel. For degree, the value '2' was chosen 10 times and for gamma the 'scale' value was chosen 6 times. These results indicate that the SVM model needed a certain flexibility to capture non-linear patterns, but did not require severe regularization.

Finally, the XGBoost model showed a strong preference for slower, more careful learning(learning_rate=0.01 or 0.1), shallow trees(max_depth=3), and stochastic regularization via subsample=0.5. The model may have benefited from simpler trees. The default value '0' for gamma was chosen 8 times. Additional split constraint may not have been necessary, since the tree depth was already limited (max_depth=3). This can indicate that the primary regularization came from the tree depth and learning rate. This combination of selected hyperparameter values suggests that the model required some complexity and careful optimization to fully capture the patterns in the dataset. Furthermore, it showed that the model had overfitting tendencies that needed multiple regularization approaches to be mitigated. There was no need for severe split constraints since tree depth already limited growth.

These findings highlight that hyperparameter sensitivity is inherently dataset-specific, with parameters such as the learning rate, kernel choice, and tree depth emerging as critical across models. While general patterns, such as conservative learning rates generalizing and performing better for the XGBoost model and the Radial Basis Function(RBF) being effective for non-linear data align with broader ML principles, the exact optimal values reflect unique dataset characteristics[71]. This underscores the importance of prioritizing tailoring tuning parameters that showed high sensitivity (e.g., XGBoost's subsample, Random Forest's split criteria), while recognizing that defaults may suffice for less impactful hyperparameters. Future work should leverage these insights by focusing tuning efforts on high-impact parameters, while balancing model-specific needs with dataset complexity.

5.4 Interpreting Feature Importance Results

This section discusses the implications of the results of the feature importance analysis across all four models. Comprehending the reason certain features were consistently ranked as important, may facilitate feature selection and engineering efforts in future work. If a specific modality or specific features, such as audio spectral characteristics, consistently exhibited high importance across models, that would indicate that those features were critical to the task.

The first model to be discussed is the Random Forest model. The top ten most important features in the Random Forest model were evenly split between text and audio modalities, suggesting a nuanced interplay between acoustic and linguistic cues in participants' descriptions of the Cookie Theft picture. Among the text features, the most prominent Empath categories were *weather*, *movement*, *sleep*, *night*, and *vacation*. Among the top ten most important features, the audio modality was represented primarily by `lspFreq` and `voiceProb`. The Random Forest model aims to make each individual branch of each tree as pure as possible, and the resulting importance scores reflect features that frequently contribute to such splits. The Cookie Theft task requires coherent and on-topic descriptions of a boy stealing cookies, a girl standing nearby, a mother washing dishes and the water overflowing from the sink. The Empath categories activated in the participants' descriptions of the Cookie Theft picture suggest that the Random Forest model is able to detect localized semantic patterns. This is mainly due to individual trees being built using different subsamples of the training data.
For the audio features, the `voiceProb` category quantifies the probability of voice activity. It reflects speech activity, capturing when participants start, stop, pause, or hesitate. In language tasks such as the Cookie Theft picture description, these aspects tie directly to fluency, hesitation, narrative flow, and coherence, which are key task-related signals[14][54]. The *voiceProb* feature helps the model separate fluent from hesitant speech, by measuring a lower value for it for hesitant speech and higher value for fluent speech. Due to fluency being a key-signal, splitting based on the *voiceProb* feature may help reduce the gini impurity, which explains the feature importance score. The *lspFreq* feature measures the distribution of energy across frequencies. So, it can effectively give an outline of a speech and distinguish certain speech, therefore, it can reduce the gini impurity, resulting in higher importance scores.

For the SVM model, the top ten most important features consisted of five text features and five audio features. The text features in order of importance were: *movement*, *weather*, *work*, *appearance*, and *clothing*. In non-linear models such as the SVM model with an RBF kernel, the feature importance is measured by how much a feature contributes to the model's ability to create separation between classes[61]. These categories may have been those that were most easily separated by a decision boundary. The *movement* category likely reflects the boy nearly falling off the stool, as indicated by the word 'tumbling' being used. The *weather* category seems to be triggered by the water falling out of the sink and flooding the room, arguably one of the most distinctive events in the picture. The *work* category was likely associated with the word 'housework', referring to the mother doing the dishes. Similarly, the *appearance* and *clothing* category may have been activated by the word 'dress', used to describe the mother's outfit. The importance scores seem to be in the order of most distinctive events to the least. The top audio features included *lspFreq*, *voiceProb* and *mfcc*. *lspFreq* values vary significantly between voices, making it very unlikely for the energy distribution among frequencies to be equal for two voices, making them useful for distinguishing between speakers. The second-ranked feature, *voiceProb*, captures voice activity. This partially overlaps with features such as *PCM Loudness*, suggesting that, while it contributes unique timing-related information, its distinctiveness is partially shared with energy-based acoustic properties such as *PCM Loudness*.

The third model, the XGBoost model, had a top ten most important features that primarily included text features, with six out of ten being text features and four audio features. The XGBoost model makes decisions by building trees sequentially, where each tree corrects the error of the previous tree. The gradient of the loss function represents how sensitive the loss function is to changes in predictions. A "high-gradient region" is an area where the model is currently making significant errors, and small improvements in prediction can lead to substantial loss reduction. The correlation matrix in figure 4 shows that the correlation between the text features is positive. The red areas in the lower right show positive correlations between text features. The *weather* category does not exist in isolation in the Cookie Theft picture description task because it represents part of a contextual framework. An example of this is the *appearance* and *weather* category, that may co-occur when describing the mother wearing a dress and the water overflowing the sink and flooding the kitchen. The predictions for the *weather* category may have produced the highest prediction errors, which in turn can make the reduction in loss relatively large when improved. Another possibility is that the *weather* category contributes the most predictive value as part of a feature combination. The audio features in the top ten most important features are dominated by *PCM Loudness*, *lspFreq*, and *voiceProb*. *PCM Loudness* might be the most important audio feature because it likely divides data into regions with similar prediction errors, making subsequent corrections more efficient. This then can lead to a steeper slope in the loss function. The second most important audio feature is the *lspFreq* category. The difference between these is marginal, approximately 0.066. The same applies to the difference between the *lspFreq* and *voiceProb* categories. This means that the reduction in loss was similar for all of them and that the various audio features likely capture related but distinct aspects of speech. No single audio feature dominated because the task required the analysis of features that capture complementary aspects of speech. Finally, the ten most important features of the baseline model were primarily text features, with seven out of ten being text features and three audio features. The baseline model is a linear regression model where the final coefficients are those that lead to the lowest mean squared error(MSE). All features were normalized prior to modeling, so importance scores reflect predictive contribution rather than differences in scale. The *domestic work* Empath category emerged as the most important feature. This may be explained by it having a relatively low correlation with other features, with its highest correlation approximating 0.5. This indicates that the information it brings is more unique. This more unique information is able to reduce the unexplained error (residuals) more, leading to a greater decrease in MSE. In contrast, the second most important category *weather* is more correlated with other features overall. As a result, it explains less unique variance and therefore contributes less to lowering the MSE. This becomes more apparent when comparing the two most important audio features with each other. The most important audio feature belongs to the *voiceProb* category and the second most important audio feature to the *lspFreq* category. The *voiceProb* category is significantly lower correlated with other features than the *lspFreq* category.

In conclusion, the analysis of the feature importance scores across the three models revealed how different machine learning models prioritized information. These models all prioritized information in distinct ways. The Random Forest model captures different localized aspects of the picture, such as semantic patterns in text (e.g., water falling on the floor belonging to the *weather* category) and speech markers (e.g., *voiceProb*) , by training each purity-driven tree on different data subsamples and aggregating the results through ensembling.
The SVM model, focused on margin maximization, emphasized features that created clear separation boundaries, particularly distinctive event-related text categories such as the boy falling off the stool related to the *movement* category, and unique spectral characteristics like *lspFreq*.
XGBoost's sequential error correction strategy uniquely favors text features as *weather* where the mistakes were similar, which created high-gradient regions where small improvements yield significant loss reduction. However, audio features showed more balanced importance due to their complementary roles in capturing different aspects of speech.
The final model, the baseline model, determines feature importance based on how much they reduce the MSE. Features that lead to a greater reduction in MSE are considered more important. The features that are more important tend to be features that have a lower correlation with other features, as they contribute more unique information to the model.
These findings highlight a key insight: feature importance scores are not an inherent property of the data.

It emerges from the interaction between the model architecture, optimization objective, and task-specific signals. This is the reason each model assigned different feature importance scores from each other. The consistent presence of both textual and audio features across all models underscores that both types of features capture complementary aspects of speech in the Cookie Theft picture description task. Future work could explore feature engineering that explicitly models the interactions to create new features that leverage the strengths of different model types for optimal performance.

5.5 Model Complexity vs. Performance

This section will discuss the trade-off between model complexity and performance, with a particular focus on whether the use of the machine learning models was justified. The baseline linear regression model performed comparably to the machine learning models, and even outperformed them in terms of the recall. The class-wise performance of the baseline model is largely indistinguishable from that of the SVM and XGBoost model. However, the ranking of the feature importance scores did differ from the machine learning models. This may be attributed to the different manners of which the scores were given across the models. The linear regression directly minimizes the loss, whereas SVM and XGBoost do so indirectly. They do so by iteratively following the gradient of the steepest descent of the loss function. Several factors may explain why the baseline model performed similarly to the more complex models. One of which may be the relatively small sample size of 674. The complex machine learning models may have been prone to slight overfitting, as the machine learning models were too complex for the small sample size. As previously mentioned in this chapter, the XGBoost model and the Random Forest model reach their AUC peak at approximately 9960 and 3404 respectively[62]. The second contributing factor for the small gap in performance of the baseline model and the machine learning models may have been the preprocessing of the features. By filtering on statistical significance and effect size, noise was reduced. Additionally, multicollinearity was reduced by removing the features that had a correlation of greater than 0.8. This makes it feasible for the linear regression model to capture patterns that may have been infeasible to capture otherwise. The more complex models may offer marginal increase in performance. This difference in performance may have been bigger if the features were either not preprocessed or preprocessed in a different manner. Finally, the number of retained features used for the models also played a role. The number of retained features was 20. Under such conditions, linear regression is able to remain stable and effective. However, if the number of retained features was equal to 200, the performance of the baseline model would have been substantially worse. This is because the machine learning models are sufficiently robust for high-dimensional data, unlike linear regression. Linear regression becomes unstable or overfits as the feature count increases, even with preprocessing, which leads to a decrease in performance[72].

Taken together, these findings allow for the addressation of whether the use of more complex machine learning models was justified in this study. The machine learning models have a higher complexity and have a higher computational cost. The performance of the machine learning models is similar to the baseline model and were even outperformed by the baseline in terms of the recall. The baseline model runs significantly faster than the machine learning models. In this case, the machine learning models achieved a similar performance to the baseline model, but at a higher computational cost. Given the absence of a clear performance advantage, the additional computational burden does not appear to be justified.

5.6 Limitations and Future Work

This section outlines the limitations of the present study and proposes directions for future research. This study has several limitations. First, the size of the dataset of 674 participants may not be sufficient to capture the full complexity of language patterns of AD patients, particularly given the acoustic diversity observed in speech patterns of Alzheimer's Disease patients. Second, while the Cookie Theft picture description task provides a standardized assessment, the test may not fully reflect naturalistic language use. Furthermore, participants may be prompted to elaborate or clarify if they omit details, introducing variability and reducing the validity of the test. As a result, the task might not generalize well to everyday conversational settings. Third, the feature extraction methods, while comprehensive, may miss subtle acoustic or linguistic patterns that could be more discriminative.

For the text features, Empath's predefined categories might miss specific linguistic patterns related to older individuals with AD. The categories such as *weather*, *sleep*, and *movement* might be too broad or not sufficiently tailored towards the capture of nuanced complex language changes associated with AD, such

as reduced syntactic complexity, pronoun overuse, or specific word-finding difficulties. The categories not being tailored towards AD detection could also make interpretation difficult, as was apparent for the water falling on the floor being put in the *weather* category. Regarding the audio features, those were extracted using the emobase feature set provided by OpenSMILE. The emobase feature set was primarily designed for emotion recognition, which may not align perfectly with detecting AD-related speech patterns[25]. AD affects language and cognition in ways that may not directly map to emotional categories. While emobase can detect a "flat tone", which is relevant for AD, it may miss signals such as word-finding difficulties. Moreover, each feature in emobase is treated with equal importance, regardless of its relevance to AD-specific vocal traits, such as word-finding pauses or disfluencies.

For future work, it is recommended to increase the sample size and increase the demographic diversity. This strengthens the models' ability to generalize. Furthermore, new feature engineering approaches should be explored, where interactions between text and audio features are explicitly modeled, potentially creating composite features that capture the complementary aspects of speech. Another promising direction is the use of a hybrid model approach. Hybrid modeling approaches can leverage the strengths of various machine learning models. As discussed in chapter 3, the Random Forest and SVM model offer robustness to overfitting and the ability to handle highly non-linear data[59]. These models may be particularly well-suited for processing textual features. The XGBoost model has the added benefit of being able to handle feature interactions without explicitly modeling them[23]. This might be more beneficial for the audio modality in this case because it has more features. The additional advantage of this is that each model can be tuned for a specific task such as tuned on the text or audio features. Smaller, specialized models for each modality are computationally less expensive than a single large model.
A practical implementation of this hybrid strategy could involve having one model, Random Forest or SVM, that captures linguistic patterns. The XGBoost model or a neural network such as a Convolutional Neural Network(CNN) could be used to capture audio patterns and extract audio features. Afterwards, the output of the text and audio models should be combined by a model such as the XGBoost model, if not used in the previous step, or a simple neural network.

In summary, while this study adds to the existing body of literature exploring multimodal approaches to Alzheimer's Disease detection, several limitations must be acknowledged. These include the modest sample size making generalization harder, task-specific constraints of the Cookie Theft picture description task, which may make it too different from natural language use, and potential misalignment of both Empath text categories and emobase audio features with AD-specific markers. Future research should seek to increase the diversity of the samples, engineer composite features that capture interaction between linguistic nuance and auditory signals, and explore a hybrid model pipeline specifically tailored to this domain. By employing modality-optimized models, including the Random Forest or SVM model for text features, XGBoost model or CNN for audio features, and unifying their outputs through a fusion model, the strengths of different models can be harnessed. Such an architecture not only improves computational efficiency but also enables robust, task-specific tuning, ultimately laying the groundwork for a scalable and adaptable Alzheimer's Disease detection tool.

6 CONCLUSION

This study investigated the research question: "How can we use machine learning to distinguish normal from disordered language use in monitoring Alzheimer's Disease in older individuals?" The findings demonstrate that machine learning, when applied to statistically selected multimodal features, effectively distinguishes disordered language associated with Alzheimer's Disease from typical language in the population of older individuals. This answer was obtained by analyzing both the transcriptions and audio segments of participants in three studies taking the Cookie Theft picture description task. For the transcripts, features were extracted by embedding each sentence of each participants' transcript using the sentence transformer 'all-mpnet-base-v2'. Cosine similarity was then computed between the sentence embeddings and the embeddings of the predefined Empath category labels. Unlike the standard Empath method, which relies on keyword matching, this approach captures higher-level semantic relationships and enables a more nuanced representation of the transcript's content. Audio features were extracted using the emobase feature set provided by OpenSMILE, which emphasizes emotion-related auditory indicators (e.g., flat tone, pauses).

The features were selected based on statistical analysis, where statistically significant($p < 0.05$) features were retained, were filtered by effect size (≥ 0.5), and checked for multicollinearity. Features with a correlation above 0.8 were removed, unless they had the highest effect size or, if equal, the highest variance. This yielded a final set of 20 features.

These features were used as input for a baseline linear regression model and three machine learning models: Random Forest, Support Vector Machine, and XGBoost. Hyperparameter tuning was conducted via 5-fold cross validation, and performance was averaged over 10 runs with different data splits. The models achieved ROC AUC scores between 0.73 and 0.76, and recall scores between 0.72 and 0.73, aligning with clinical benchmarks. The results directly address the research question: machine learning models successfully leveraged the selected features to distinguish ordered from disordered language in older individuals. This study also revealed unexpected insights: audio features contributed comparably to text features, challenging prior assumptions that text features dominate Alzheimer's Disease classification. This finding underscores the complementary nature of multimodal data in capturing nuanced language and acoustic patterns associated with cognitive decline. Additionally, the similar performance of all three machine learning models compared to each other and the baseline model suggests that the core signal statistically selected features were the primary drivers of performance, rather than algorithmic complexity. This raises the question of whether the added computational cost of more complex machine learning models was justified in this case. Given that the baseline linear regression achieved comparable results while being computationally more efficient, the use of advanced models may not offer substantial practical advantages under the current conditions. This outcome could also have been a reflection of the limited sample size, which may have constrained the Random Forest and SVM models from realizing their full potential.

While the study answers the research question affirmatively, limitations remain. The dataset size of 674 participants limits generalizability. Furthermore, the task-specific nature of the Cookie Theft picture description task may not reflect naturalistic conversational contexts, which may also hinder generalizability. Additionally, feature extraction tools used in the present study, Empath and OpenSMILE, may not fully capture the nuanced linguistic or acoustic patterns related to Alzheimer's Disease, such as word-finding difficulties or syntactic complexity. Future work should focus on using an expanded dataset, engineering Alzheimer's Disease-specific features, and exploring hybrid model approaches to refine detection accuracy. These directions could enhance model performance and support the development of scalable, multimodal tools for monitoring Alzheimer's Disease in older individuals in real-world clinical and home-based settings.

REFERENCES

[1] (ADI), A. D. I. (n.d.). Adi - dementia statistics.

[2] Ahlbom, A. and Bottai, M. (2016). False-positive findings, multiple comparisons and the strength of hypotheses. *Journal of Internal Medicine*, 279(4):399–402.

[3] Al-Hammadi, M., Fleyeh, H., Åberg, A. C., Halvorsen, K., and Thomas, I. (2024). Machine learning approaches for dementia detection through speech and gait analysis: A systematic literature review. *Journal of Alzheimer's Disease*, 100(1):1–27.

[4] Alatrany, A. S., Khan, W., Hussain, A., Kolivand, H., and Al-Jumeily, D. (2024). An explainable machine learning approach for alzheimer's disease classification. *Scientific reports*, 14(1):2637.

[5] Alizamir, M., Wang, M., Ikram, R. M. A., Gholampour, A., Ahmed, K. O., Heddam, S., and Kim, S. (2025). An interpretable xgboost-shap machine learning model for reliable prediction of mechanical properties in waste foundry sand-based eco-friendly concrete. *Results in Engineering*, page 104307.

[6] Association, A. S.-L.-H. et al. (1993). Definitions of communication disorders and variations.

[7] audEERING GmbH (2025). *openSMILE Documentation*. Accessed: 2025-05-11.

[8] Auxier, B. and Anderson, M. (2021). Social media use in 2021.

[9] Becker, J. T., Boller, F., Lopez, O. L., Saxton, J., and McGonigle, K. L. (1994). The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594.

[10] Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

[11] Burke, E., Gunstad, J., Pavlenko, O., and Hamrick, P. (2024). Distinguishable features of spontaneous speech in alzheimer's clinical syndrome and healthy controls. *Aging, Neuropsychology, and Cognition*, 31(3):575–586.

[12] Chou, C.-J., Chang, C.-T., Chang, Y.-N., Lee, C.-Y., Chuang, Y.-F., Chiu, Y.-L., Liang, W.-L., Fan, Y.-M., and Liu, Y.-C. (2024). Screening for early alzheimer's disease: enhancing diagnosis with linguistic features and biomarkers. *Frontiers in Aging Neuroscience*, 16:1451326.

[13] Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. routledge.

[14] Cummings, L. (2019). Describing the cookie theft picture: Sources of breakdown in alzheimer's dementia. *Pragmatics and Society*, 10(2):153–176.

[15] Data Science Council of America (DASCA) (2024). Understanding real-time data analytics and how it works. Accessed: 2025-05-11.

[16] Doğdu, C., Kessler, T., Schneider, D., Shadaydeh, M., and Schweinberger, S. R. (2022). A comparison of machine learning algorithms and feature sets for automatic vocal emotion recognition in speech. *Sensors*, 22(19):7561.

[17] Drougkas, G., Bakker, E. M., and Spruit, M. (2024). Multimodal machine learning for language and speech markers identification in mental health. *BMC Medical Informatics and Decision Making*, 24(1):354.

[18] Endalie, D. and Abebe, W. T. (2023). Analysis of lung cancer risk factors from medical records in ethiopia using machine learning. *PLOS Digital Health*, 2(7):e0000308.

[19] Eyben, F., Wöllmer, M., and Schuller, B. (2010). opensmile – the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 1459–1462. ACM.

[20] Eyigoz, E., Mathur, S., Santamaria, M., Cecchi, G., and Naylor, M. (2020). Linguistic markers predict onset of alzheimer's disease. *EClinicalMedicine*, 28.

[21] Fast, E., Chen, B., and Bernstein, M. S. (2016). Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4647–4657.

[22] Ghasemzadeh, H., Hillman, R. E., and Mehta, D. D. (2024). Toward generalizable machine learning models in speech, language, and hearing sciences: Estimating sample size and reducing overfitting. *Journal of Speech, Language, and Hearing Research*, 67(3):753–781.

[23] Goyal, K., Dumancic, S., and Blockeel, H. (2020). Feature interactions in xgboost. *arXiv preprint arXiv:2007.05758*.

[24] Gustavsson, A., Norton, N., Fast, T., Frölich, L., Georges, J., Holzapfel, D., Kirabali, T., Krolak-Salmon, P., Rossini, P. M., Ferretti, M. T., et al. (2023). Global estimates on the number of persons across the alzheimer's disease continuum. *Alzheimer's & Dementia*, 19(2):658–670.

[25] Haider, F., Pollak, S., Albert, P., and Luz, S. (2019). Extracting audio-visual features for emo-

tion recognition through active feature selection. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1–5. IEEE.

[26] Hammers, D. B., Eloyan, A., Taurone, A., Thangarajah, M., Beckett, L., Gao, S., Kirby, K., Aisen, P., Dage, J. L., Foroud, T., et al. (2023). Profiling baseline performance on the longitudinal early-onset alzheimer's disease study (leads) cohort near the midpoint of data collection. *Alzheimer's & Dementia*, 19:S8–S18.

[27] Hason, L. and Krishnan, S. (2022). Spontaneous speech feature analysis for alzheimer's disease screening using a random forest classifier. *Frontiers in Digital Health*, 4:901419.

[28] Hefner, J. T. and Ousley, S. D. (2014). Statistical classification methods for estimating ancestry using morphoscopic traits. *Journal of Forensic Sciences*, 59(4):883–890.

[29] Hernández-Domínguez, L., Ratté, S., Sierra-Martínez, G., and Roche-Bergua, A. (2018). Computer-based evaluation of alzheimer's disease and mild cognitive impairment patients during a picture description task. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:260–268.

[30] Hofmann, T., Schölkopf, B., and Smola, A. J. (2008). Kernel methods in machine learning.

[31] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

[32] Huang, L., Yang, H., Che, Y., and Yang, J. (2024). Automatic speech analysis for detecting cognitive decline of older adults. *Frontiers in Public Health*, 12:1417966.

[33] Ismail, M. S., Brand, C., and Martin, K. (2007). Benefits of early pharmacological treatment in alzheimer disease. *Psychiatric Times*, 24(4):49–49.

[34] James, C., Ranson, J. M., Everson, R., and Llewellyn, D. J. (2021). Performance of machine learning algorithms for predicting progression to dementia in memory clinic patients. *JAMA network open*, 4(12):e2136553–e2136553.

[35] Jiang, F., Dong, Q., Wu, S., Liu, X., Dayimu, A., Liu, Y., Ji, H., Wang, L., Liu, T., Li, N., et al. (2024). A comprehensive evaluation on the associations between hearing and vision impairments and risk of all-cause and cause-specific dementia: results from cohort study, meta-analysis and mendelian randomization study. *BMC medicine*, 22(1):518.

[36] Kaiser, S., Lyne, J., Agartz, I., Clarke, M., Mørch-Johnsen, L., and Faerden, A. (2017). Individual negative symptoms and domains–relevance for assessment, pathomechanisms and treatment. *Schizophrenia research*, 186:39–45.

[37] Kaiying, L. and PW, Y. (2024). Multimodal deep learning for dementia classification using text and audio [j]. *Scientific Reports*, 14(1):13887–13887.

[38] Li, J., Li, Y., Wang, Y., Wu, X., and Meng, H. (2024). Devising a set of compact and explainable spoken language feature for screening alzheimer's disease. In *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 471–475. IEEE.

[39] Lin, H., Karjadi, C., Ang, T. F., Prajakta, J., McManus, C., Alhanai, T. W., Glass, J., and Au, R. (2020). Identification of digital voice biomarkers for cognitive health. *Exploration of medicine*, 1:406.

[40] Louppe, G. (2014). *Understanding random forests: From theory to practice*. PhD thesis, Universite de Liege (Belgium).

[41] Lu, M. (2023). Dementiabank english lu corpus. https://dementia.talkbank.org/access/English/Lu.html. DOI: https://doi.org/10.21415/4KE0-6348.

[42] Lyons, M., Aksayli, N. D., and Brewer, G. (2018). Mental distress and language use: Linguistic analysis of discussion forum posts. *Computers in Human Behavior*, 87:207–211.

[43] Magnin, B., Mesrob, L., Kinkingnéhun, S., Pélégrini-Issac, M., Colliot, O., Sarazin, M., Dubois, B., Lehéricy, S., and Benali, H. (2009). Support vector machine-based classification of alzheimer's disease from whole-brain anatomical mri. *Neuroradiology*, 51:73–83.

[44] Mahon, E. and Lachman, M. E. (2022). Voice biomarkers as indicators of cognitive changes in middle and later adulthood. *Neurobiology of aging*, 119:22–35.

[45] Mei, A. L. (2024). How does language influence our minds? from a linguistics perspective. *Lecture Notes in Education Psychology and Public Media*, 42:205–209.

[46] Miller, V. A., Erlien, S., and Piersol, J. (2014). Support vector machine classification of dimensionally reduced structural mri images for dementia. *arXiv preprint arXiv:1406.6568*.

[47] Newmark, J., Gebara, M. A., Aizenstein, H., and Karp, J. F. (2020). Engaging in late-life mental health research: a narrative review of challenges to participation. *Current Treatment Options in Psychiatry*, 7:317–336.

[48] Niemelä, M., von Bonsdorff, M., Äyrämö, S., and Kärkkäinen, T. (2025). Dementia classification using acoustic speech and feature selection. *arXiv preprint arXiv:2502.03484*.

[49] NVIDIA (n.d.). What is xgboost?

[50] Ocampo Osorio, F., Alzate-Ricaurte, S., Mejia Vallecilla, T. E., and Cruz-Suarez, G. A. (2024). The anesthesiologist's guide to critically assessing machine learning research: a narrative review. *BMC anesthesiology*, 24(1):452.

[51] Organization, P. . P. A. H. (2013). Depression and dementia are top mental disorders for people over 60.

[52] Organization, W. H. (2022). Who highlights urgent need to transform mental health and mental health care.

[53] Organization, W. H. (n.d.). Mental health of older adults.

[54] Ossewaarde, R. (2025). Automated measurements of fluency, syntax and semantics in the language of persons with primary progressive aphasia.

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

[56] Pfisterer, F., Van Rijn, J. N., Probst, P., Müller, A. C., and Bischl, B. (2021). Learning multiple defaults for machine learning algorithms. In *Proceedings of the genetic and evolutionary computation conference companion*, pages 241–242.

[57] Priyadarshinee, P., Clarke, C. J., Melechovsky, J., Lin, C. M. Y., BT, B., and Chen, J.-M. (2023). Alzheimer's dementia speech (audio vs. text): Multi-modal machine learning at high vs. low resolution. *Applied Sciences*, 13(7):4244.

[58] Saeedi, S., Hetjens, S., Grimm, M., and Latoszek, B. B. v. (2024). Acoustic speech analysis in alzheimer's disease: A systematic review and meta-analysis. *The Journal of Prevention of Alzheimer's Disease*, 11(6):1789–1797.

[59] Sarica, A., Cerasa, A., and Quattrone, A. (2017). Random forest algorithm for the classification of neuroimaging data in alzheimer's disease: a systematic review. *Frontiers in aging neuroscience*, 9:329.

[60] Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

[61] scikit learn (n.d.).

[62] Silvey, S. and Liu, J. (2024). Sample size requirements for popular classification algorithms in tabular clinical data: Empirical study. *Journal of Medical Internet Research*, 26:e60231.

[63] Sørensen, L., Nielsen, M., Initiative, A. D. N., et al. (2018). Ensemble support vector machine classification of dementia using structural mri and mini-mental state examination. *Journal of neuroscience methods*, 302:66–74.

[64] Spruit, M., Verkleij, S., de Schepper, K., and Scheepers, F. (2022). Exploring language markers of mental health in psychiatric stories. *Applied Sciences*, 12(4):2179.

[65] Tay, R. (2017). Correlation, variance inflation and multicollinearity in regression model. *Journal of the Eastern Asia Society for Transportation Studies*, 12:2006–2015.

[66] Twait, E. L., Andaur Navarro, C. L., Gudnason, V., Hu, Y.-H., Launer, L. J., and Geerlings, M. I. (2023). Dementia prediction in the general population using clinically accessible variables: a proof-of-concept study using machine learning. the ages-reykjavik study. *BMC medical informatics and decision making*, 23(1):168.

[67] U.S.News (2024). $282 billion: What mental illness costs america each year.

[68] Wei, H. T., Kulzhabayeva, D., Erceg, L., Robin, J., Hu, Y. Z., Chignell, M., and Meltzer, J. A. (2024). Cognitive components of aging-related increase in word-finding difficulty. *Aging, Neuropsychology, and Cognition*, 31(6):987–1019.

[69] Whelan, R., Barbey, F. M., Cominetti, M. R., Gillan, C. M., and Rosická, A. M. (2022). Developments in scalable strategies for detecting early markers of cognitive decline. *Translational Psychiatry*, 12(1):473.

[70] XGBoost (n.d.). Feature interaction constraints.

[71] XGBoosting (n.d.). Tune xgboost "learning_rate" parameter.

[72] Yoo, W., Mayberry, R., Bae, S., Singh, K., He, Q. P., and Lillard Jr, J. W. (2014). A study of effects of multicollinearity in the multivariable analysis. *International journal of applied science and technology*,

4(5):9.