Universiteit Leiden
The Netherlands

Bachelor Data Science & Artificial Intelligence

Assessing Bias in Machine Learning Models for Alzheimer's Disease Detection Across Gender and Age

Rebeca Sanz Lozano
s3539911

Supervisors:
Prof.dr. M.R. Spruit & Dr. B.M.A. van Dijk

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

01/07/2025

Abstract

Early detection of Alzheimer's Disease (AD) is essential because available treatments are more effective in the early stages of the disease, potentially slowing cognitive decline and helping patients maintain independence for longer. However, current diagnostic tools are often costly, invasive, or inaccessible. This thesis explores the use of transcribed speech for automated AD detection using two approaches: a classical Random Forest (RF) model with engineered linguistic features, and a transformer-based RoBERTa model using raw text. Using the Pitt Corpus and the ADReSS datasets, the performance, interpretability, and fairness of both models across age and gender groups were evaluated. The results show that both models perform similarly on the ADReSS dataset (accuracy for both: 0.833; ROC-AUC: 0.889 for RF, 0.901 for RoBERTa), while RoBERTa outperforms RF on the Pitt dataset (accuracy: 0.794 vs 0.720; ROC-AUC: 0.876 for RoBERTa, 0.800 for RF). However, the Random Forest model provides greater transparency. Fairness analysis revealed demographic disparities in both models, particularly related to age. RoBERTa showed lower bias on the ADReSS dataset for the age group, but more bias than RF on the Pitt dataset. Post-processing bias mitigation techniques, such as Reject Option Classification (ROC), significantly improved fairness metrics, without substantial reductions in performance. For example, the Disparate Impact (DI) metric for the age group in the ADReSS dataset using Random Forest was reduced from 1.667 to the ideal value 1.000, while accuracy remained unchanged at 0.833. In some cases, accuracy even improved slightly after mitigation. The results emphasize that fairness and interpretability are as crucial as predictive accuracy when applying AI in healthcare.

Contents

Abbreviations

Medical and Clinical Terms

AD	Alzheimer's Disease
MCI	Mild Cognitive Impairment
MMSE	Mini-Mental State Examination
HC	Healthy Control

Machine Learning and NLP Models

AI	Artificial Intelligence
ML	Machine Learning
NLP	Natural Language Processing
RF	Random Forest
BERT	Bidirectional Encoder Representations from Transformers
LLMs	Large Language Models
ANN	Artificial Neural Network
KNN	K-Nearest Neighbors
SVM	Support Vector Machine

Datasets and Tools

ADReSS	Alzheimer's Dementia Recognition through Spontaneous Speech
AIF360	AI Fairness 360
CHILDES	Child Language Data Exchange System
CHAT	Codes for the Human Analysis of Transcripts
ACL	Association for Computational Linguistics

Explainability and Fairness Metrics

SHAP	SHapley Additive exPlanations
SPD	Statistical Parity Difference
EOD	Equal Opportunity Difference
AOD	Average Odds Difference
DI	Disparate Impact
AUC	Area Under the Curve

Linguistic and Transcript Analysis

IU	Information Unit
CIU	Correct Information Unit
PoS	Part of Speech

1 Introduction

Dementia is a commonly used term to describe a variety of symptoms that include loss of memory, difficulties with problem-solving skills and other thinking abilities, cognitive decline in general and language problems, among others. Between 60% to 80% of dementia cases are caused by Alzheimer's Disease (AD) (Alzheimer's Association, 2024). AD is a neurodegenerative condition characterized by the accumulation of beta-amyloid plaques and tau tangles in the brain. This results in synaptic dysfunction, leading to neuronal death and an associated inflammatory response (de Oliveira et al., 2021). Clinically, AD manifests as a gradual decline in cognitive abilities, ultimately impacting an individual's ability to carry out daily activities. AD can enter the preclinical phase up to 20 years before the appearance of symptoms, after which the condition progresses into the mild cognitive impairment (MCI) phase (Rafii & Aisen, 2023). Early detection is crucial because, although there is no cure, treatments are generally more effective during the MCI phase, potentially slowing the progression of the disease (Chen & Wang, 2013).

As Park et al. (2023) explains, current diagnostic protocols follow a multimodal approach that includes neuropsychological testing, neuroimaging (e.g., MRI or PET), and cerebrospinal fluid (CSF) biomarkers. Although these methods are useful, they can be expensive, invasive, and not widely accessible. This makes it even more important to find more scalable diagnostic tools. A potential approach involves using language as an early indicator of AD. Artificial Intelligence (AI) and machine learning (ML) techniques are increasingly being used in healthcare to improve diagnostics and decision-making (Alanazi, 2022). Recent progress in the fields of AI and Natural Language Processing (NLP) have resulted in the creation of non-invasive, automated techniques for detecting AD using linguistic features from speech data.

Studies have consistently shown that individuals diagnosed with AD exhibit specific language impairments, especially in lexical and semantic processing, syntactic structure, and speech fluency. Deficits in lexical and semantic abilities are defined by a reduced vocabulary, an increase in the frequency of word repetition, and a greater dependence on pronouns. These symptoms imply impairments in lexical retrieval and word selection, which is known as anomia (Kavé & Goral, 2016). Syntactically, AD patients generally produce shorter and simpler grammatically sentences, which limits their communication abilities (Ortiz et al., 2024). Furthermore, language impairments are exacerbated by speech disturbances such as speech pauses, hesitations, and word-finding difficulties, all affecting general verbal communication (Olmos-Villaseñor et al., 2023). Language impairments reflect the decline in semantic memory and executive function, and research suggests that they may appear even before a formal diagnosis has been made (Verma & Howard, 2012).

Transformer-based deep learning models have significantly improved AD classification from speech and text data. Through textual transcriptions of speech samples and the linguistic markers associated with AD, these models have shown impressive performance distinguishing AD patients from healthy controls. Recent studies have built upon this progress by showing the efficacy of specific transformer architectures, including BERT (Bidirectional Encoder Representations from Transformers). Instead of depending on manually extracted linguistic features, BERT models are trained directly using raw text data. In research conducted by Balagopalan et al. (2021), fine-tuned BERT models obtained an accuracy of 83.3% on unseen test data, demonstrating their ability

to detect changes in language in AD patients with minimal feature engineering. However, these deep learning models generally need significant computational resources and large datasets for fine-tuning, which might limit their practical implementation in some healthcare environments.

To address these considerations, this thesis uses two approaches to detect AD using language data: a classical machine learning model known as Random Forest (RF) and a transformer-based deep learning model called RoBERTa. RF is a popular ML model applied in healthcare settings due to its robustness, interpretability, and effectiveness with smaller datasets. Furthermore, they offer a more accessible solution to those institutions with limited computational resources. On the other hand, RoBERTa is a more recent and robustly optimized variant of BERT that has achieved good results in many NLP tasks (Y. Liu et al., 2019). It can identify subtle linguistic impairments in both semantic and syntactic structures.

1.1 Problem statement

Previous studies have shown the potential of machine learning models in detecting AD, but the issue of fairness within these AI systems has not been thoroughly investigated. Many studies operate under the assumption that balancing datasets is sufficient to address bias; however, biases can still arise at the model level, in particular in the interpretation of linguistic features. This thesis aims to explore whether machine learning models for AD detection reveal disparities based on gender and age groups, as well as to evaluate how effectively fairness-aware AI strategies can mitigate these biases without compromising classification accuracy.

In the healthcare context, biased outputs could have severe implications, such as unequal access to diagnosis and treatment. From a personal perspective, I believe it is especially important to ensure that less frequently diagnosed groups, such as males under 65, are not overlooked. Although these individuals may not fit the typical demographic profile of AD cases, their diagnostic needs are equally important. Ensuring fair model behavior across all subgroups is crucial to building equitable and trustworthy AI systems in healthcare. To explore this gap, this thesis aims to answer the following research questions:

- How do gender and age influence the predictive performance of machine learning models in Alzheimer's Disease detection?
- Can fairness-aware AI techniques mitigate these biases while maintaining classification accuracy?

To address the research questions, the following approach was adopted. First, a systematic literature review was conducted to identify state-of-the-art models for Alzheimer's Disease detection. The goal was to choose one traditional machine learning model and one transformer-based model that showed strong performance. Including a machine learning model was motivated by its reduced computational costs and to assess whether competitive results could still be achieved when compared to more complex architectures. Once the models were selected, they were trained and evaluated on both datasets used in this thesis. In addition to measuring overall performance, explainability was also considered a key aspect, particularly in the healthcare sector. Therefore, SHAP (SHapley

Additive exPlanations) was applied for both models to provide an understanding of the decision making process. Finally, fairness metrics were computed, and fairness-aware mitigation techniques were implemented to evaluate their impact on both equity and model performance.

1.2 Thesis overview

This thesis is divided into different chapters. Abbreviations provides a list of abbreviations used throughout the thesis. Section 1 introduces the research context, defines the problem, and outlines the motivation and objectives of the thesis. Section 2 contains the literature review, including the screening methodology, a discussion of selected studies on AD detection, and a discussion of algorithmic bias and fairness in healthcare. Section 3 details the methodology pipeline; covering the datasets used, the preprocessing steps, and the extraction of linguistic features relevant to AD. It also describes the model setup for the Random Forest and RoBERTa approaches, as well as the fairness techniques used, including the metrics and the mitigation. Section 4 presents the results of the experiments, including model performance, explainability analysis using SHAP, and fairness evaluation across gender and age. Finally, Section 5 summarizes the thesis's contributions and provides directions for future research.

As part of the bachelor program in Data Science and Artificial Intelligence at the Leiden Institute of Advanced Computer Science (LIACS) of Leiden University, this thesis was written under the supervision of Prof.dr. M.R. Spruit and Dr. B.M.A. van Dijk.

2 Literature Review

2.1 Screening Procedure

A systematic literature review was conducted in accordance with the SYMBALS framework introduced by Van Haastrecht et al. (2021). Given the rapid advancements in ML and NLP techniques for AD detection, especially using Large Language Models (LLMs), the backward snowballing phase was omitted to focus only on current relevant literature.

Two sources were used to search for relevant publications: Association for Computational Linguistics (ACL) Anthology and Europe PMC. From the ACL Anthology, all papers from the main ACL conferences held in 2024 and 2025 were included, resulting in a collection of 4395 papers. To complement this with papers from the medical field, Europe PMC was queried with a targeted search query, filtering for papers published between January 2020 and March 2025. This resulted in an additional set of 1167 papers. After merging the datasets and removing duplicates, the final corpus consisted of 5279 unique entries, which were exported into a single CSV file for further processing.

The search query used for PMC was the following:

```
(("Alzheimer's Disease" OR Dementia OR "Cognitive Decline")
 AND
("Linguistic Markers" OR "Lexical Features" OR "Text Analysis"
 OR "Speech Processing" OR "Word embeddings" OR "Language models"
 OR "Syntactic analysis" OR "Semantic analysis")
 AND
("Machine Learning" OR "Deep Learning" OR "Natural Language Processing"
 OR "NLP" OR "Transformer" OR "BERT" OR "Artificial Intelligence"
 OR "AI" OR "Neural Networks"))
```

The first stage of the screening process was performed using ASReview, an active learning tool. The model was first initialized with a limited number of selected examples, consisting on five relevant and five irrelevant papers. The papers were identified using the random selection tool available in ASReview, which takes random samples from the full dataset. This approach was chosen to minimize possible selection bias at the beginning stage to guarantee an impartial starting point for the model. ASReview was configured with its default settings, which are as follows:

```
Feature extraction technique: TF-IDF
Classifier: Naive Bayes
Query strategy: Maximum
Balance strategy: Dynamic resampling (Double)
```

In order to determine when to stop the screening process, a stopping criterion was defined: if 20 papers in a row were found to be irrelevant, the process would be stopped. This criterion was met after screening 210 papers (3.98% of the total dataset), out of which 54 were identified relevant to the study. Therefore, 156 papers were excluded during this phase. Figure 1 illustrates the overall review process.

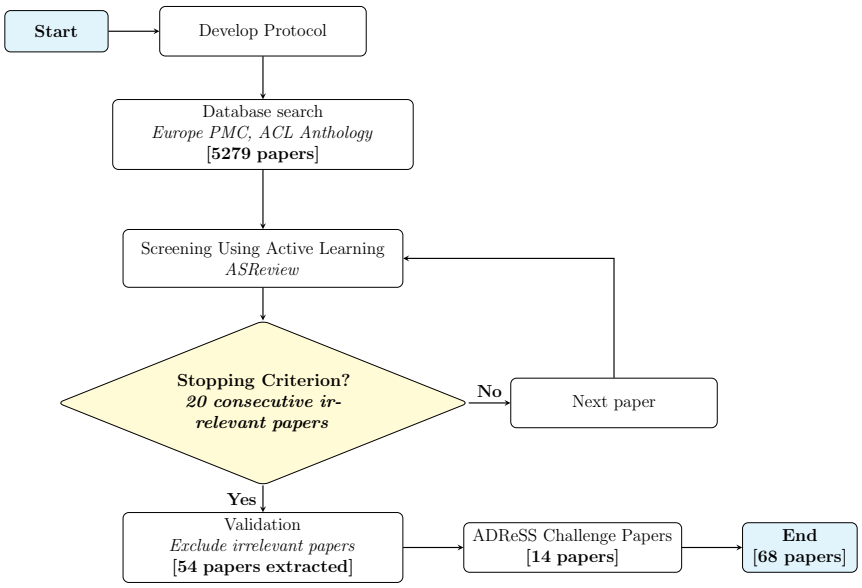


Figure 1: Flowchart of the Literature Review Process

Table 1 lists the initial set of papers used to initiate ASReview. It includes both relevant and irrelevant examples, along with a short justification for each.

Table 1: Papers initially labeled as relevant and irrelevant for training the ASReview model.

Paper	Relevance	Reason
Lindsay et al. (2021)	Relevant	Multilingual NLP for AD speech detection with explainable ML.
Parsapoor et al. (2023)	Relevant	Evaluates how input factors affect dementia ML performance.
Wegner et al. (2024)	Relevant	Tool for harmonizing AD datasets to enable cross-cohort studies.
Mao et al. (2023)	Relevant	Uses BERT on clinical notes to predict MCI-to-AD progression.
Fu et al. (2024)	Relevant	Combines acoustic and semantic speech features for AD classification.
Das et al. (2024)	Irrelevant	Focuses on mathematical reasoning using LLMs.
Ren et al. (2024)	Irrelevant	Discusses watermarking for generated text.
Wu et al. (2024)	Irrelevant	Targets neural machine translation.
Deng et al. (2024)	Irrelevant	Focus on attribution in summarization.
X. Zhang et al. (2024)	Irrelevant	Works on emotion recognition in domain adaptation.

In addition to the reviewed literature, the full collection of 14 papers from the ADReSS Challenge was included in the review. These papers were considered essential to provide a comprehensive view of the field, as they are widely recognized as benchmarks in research on language-based approaches to AD detection.

2.2 Selected Literature and Models

2.2.1 Relevant Papers

This section highlights five particularly relevant studies, selected from the 54 papers reviewed, that are most closely aligned with the goals of this thesis. The primary focus was on models that process transcribed speech rather than acoustic signals, since this thesis focuses only on transcripts. Each selected paper demonstrates methodological innovation, strong predictive results, or architectural relevance to this research.

AD-BERT: Using pre-trained language model to predict the progression from mild cognitive impairment to Alzheimer's disease

Mao et al. (2023) present AD-BERT, a deep learning framework based on a fine-tuned Bio+Clinical BERT model, developed to predict progression from MCI to AD using unstructured clinical notes from electronic health records. Although the authors use written clinical documentation rather than spontaneous speech, their approach is relevant to this research because it also relies on language-based features and transformer models to assess cognitive decline. The model achieves strong results, with an AUC of 0.883 and F1-score of 0.68 when evaluated on data from an external healthcare institution. An important distinction is that the authors focus on predicting disease progression, whereas in this thesis the focus is on AD detection or diagnosis using spontaneous speech. Nevertheless, both studies highlight the value of natural language representations in modeling cognitive decline. The study also shows that domain-specific adaptation of pre-trained models can improve performance, which supports the use of BERT for speech-based classification tasks.

Improving Alzheimer's Disease Detection for Speech Based on Feature Purification Network

N. Liu, Yuan, and Tang (2022) present a study on AD detection using transcribed speech, introducing a novel deep learning architecture called GP-Net, which aims to improve the performance of transformer models by enhancing the quality of the extracted features. The method works by filtering out linguistic patterns that are common to both groups, and thus, not helpful for classification, allowing the model to focus on the textual features that are more specific to each group and actually help distinguish them.
The model consists of two networks: a common feature extractor (G-Net), which uses a gradient reversal layer to identify the shared features, and a purification network (P-Net), which removes these features from the overall set. The authors apply this methodology to three datasets, including DementiaBank Pitt Corpus and the ADReSS dataset, which are the same datasets used in this study. The model achieves an accuracy of 0.935 and a F1-score of 0.912, outperforming baseline BERT and traditional classifiers.
A limitation of the study is that it primarily compares deep learning methods without analyzing how purified features align with known clinical or linguistic markers of AD. Additionally, the interpretability of the purified features remains limited due to the complexity of the model.

Revealing the Roles of Part-of-Speech Taggers in Alzheimer Disease Detection: Scientific Discovery Using One-Intervention Causal Explanation

Wen et al. (2023) investigate the role of part-of-speech (PoS) features in AD detection using

transformer-based model trained on transcripts from the DementiaBank Pitt Corpus. Their goal is to understand whether individual speech features are useful for diagnosis and to identify the most predictive ones.The model uses counts of 27 PoS features and achieves high performance, with an accuracy of 0.922 and a F1-score of 0.955.
To interpret the predictions, the authors introduce a novel explainability method called one-intervention causal explanation (OICE). This method evaluates the importance of each feature by simulating changes to it while taking into account causal relationships between the features. The authors identified 12 PoS features that are particularly influential in classification, these include: personal pronouns, gerund verbs, common nouns and adverbs.
A limitation of the study is that it only focuses on PoS features and does not include other linguistic information. Although the use of causal inference helps explain why the model makes some decisions, the reliability depends on how accurately the model has captured the true causal relationships between language features. If they are incorrect, the explanations may not reflect the true reasons for the predictions.

Semantic Feature Extraction Using SBERT for Dementia Detection

Santander-Cruz et al. (2022) propose a methodology for detecting dementia using semantic, syntactic, and lexical features extracted from spontaneous speech transcripts. Their approach focuses on generating sentence embeddings using the Sentence-BERT (SBERT) model and combining these features with classifiers like support vector machines (SVM), random forest, k-nearest neighbors (KNN), and artificial neural networks (ANN). The authors evaluate their method on 550 narrative speech samples from the DementiaBank Pitt Corpus, mainly from the Cookie Theft picture description task. They show that adding semantic features improves performance over using only lexical and syntactic features. Their best results were obtained using SVM with a polynomial kernel, with an accuracy of 0.77 and an F1-score of 0.80. A limitation of the study is the limited interpretability of the semantic embeddings and the lack of exploration of how these features align with established clinical or linguistic markers.

A Transfer Learning Method for Detecting Alzheimer's Disease Based on Speech and Natural Language Processing

N. Liu, Luo, et al. (2022) present a study on AD detection using spontaneous speech and transfer learning, applying a pre-trained DistilBERT language model combined with a logistic regression classifier. The authors evaluate their method on the ADReSS dataset which achieves an accuracy of 0.88. They compare different classifiers and conclude that logistic regression performs best when paired with DistilBERT embeddings. A limitation of the study is the interpretability, as deep learning models with many parameters make it difficult to understand why a prediction was made.

2.2.2 Algorithmic Bias and Fairness

The growing use of ML in sensitive areas such as healthcare has raised important concerns about algorithmic bias and fairness, especially when it comes to diagnostic tasks that involve differences in language or demographics. In this thesis, algorithmic bias is defined as systematic disparities in model performance or behavior that result from spurious correlations, confounding variables, or

data imbalances, rather than reflecting the differences in clinical conditions. Therefore, fairness refers to the goal of minimizing those disparities, especially if they may disadvantage particular subgroups. While many fairness definitions and strategies have been proposed in the literature, integrating them into clinical workflows is difficult due to noisy data, small datasets, and other influencing factors (Pessach & Shmueli, 2023).

In the context of AD detection, research on fairness is still relatively limited. One notable study by Y.-L. Liu et al. (2024) revealed a Clever Hans Effect in AD classification models trained on the Pitt Corpus. They discovered that these models could achieve nearly perfect accuracy by just analyzing the silent parts of the audio recordings, which, theoretically, should not contain meaningful diagnostic information. This unexpected performance was caused by environmental noise and other recording artifacts that coincidentally matched the diagnostic labels. The study shows how easily models can pick up on irrelevant patterns in the data, which can make their results seem better than they actually are. Although the study does not focus on underrepresented groups, it is relevant to algorithmic fairness as it exemplifies how models can learn spurious patterns.

A more specific example of fairness in AD detection is provided by Zhu et al. (2019), who demonstrated that age, which is a natural associated factor with dementia, acts as a confounding variable in ML models trained in linguistic features. The study introduced fair representation learning models that separate age from the features used in the classification. These models preserved most of the predictive power while significantly reducing age-related bias.

Apart from AD, broader healthcare fairness studies have emphasized similar concerns. Mosteiro et al. (2022) studied fairness in mental health classification models trained on psychiatric records. They found gender disparities in model performance, which they were able to partially fix using reweighting and discrimination-aware regulations. These steps reduced bias without negatively impacting accuracy.

To support fairness detection and mitigation, several software libraries have been developed. Both AI Fairness 360 (AIF360) (Bellamy et al., 2019) and Fairlearn (Bird et al., 2020) are widely used fairness evaluation libraries. While Fairlearn focuses on interpretable mitigation techniques for classification and regression tasks, AIF360 offers a more extensive set of tools for bias detection metrics and algorithmic mitigation, including pre-, in- and post-processing methods (Pandey, 2022). Due to its richer selection of healthcare relevant bias metrics and algorithms, AIF360 is the preferred choice for this thesis.

Finally, W. Zhang (2024) argues that much of the AI fairness literature is developed under idealized assumptions, such as clean, independently, and identically distributed (IID) data with static labels, which rarely occur in real healthcare settings. In clinical NLP, data is often messy, incomplete, and highly variable across different populations, presenting considerable challenges for the application of standard techniques aimed at reducing bias. These examples illustrate that fairness in clinical ML cannot be taken for granted. Therefore, it is crucial to evaluate models not only for accuracy but also for fairness.

3 Method

This section describes the methodology used to investigate AD detection using Random Forest and RoBERTa models. Figure 2 provides an overview of the complete experimental pipeline. Section 3.1 describes the datasets used in the thesis, including their origin and characteristics. Section 3.2 explains the preprocessing steps applied to the data, including cleaning and feature extraction. Section 3.3 outlines the linguistic features on which Random Forest is based and how the features were extracted. Section 3.4 describes the model setup and includes details on the Random Forest and RoBERTa models used for classification. Finally, Section 3.5 discusses the fairness-aware techniques applied. The full implementation of the models and experiments can be found at: https://github.com/beckiechill/bachelor-thesis.

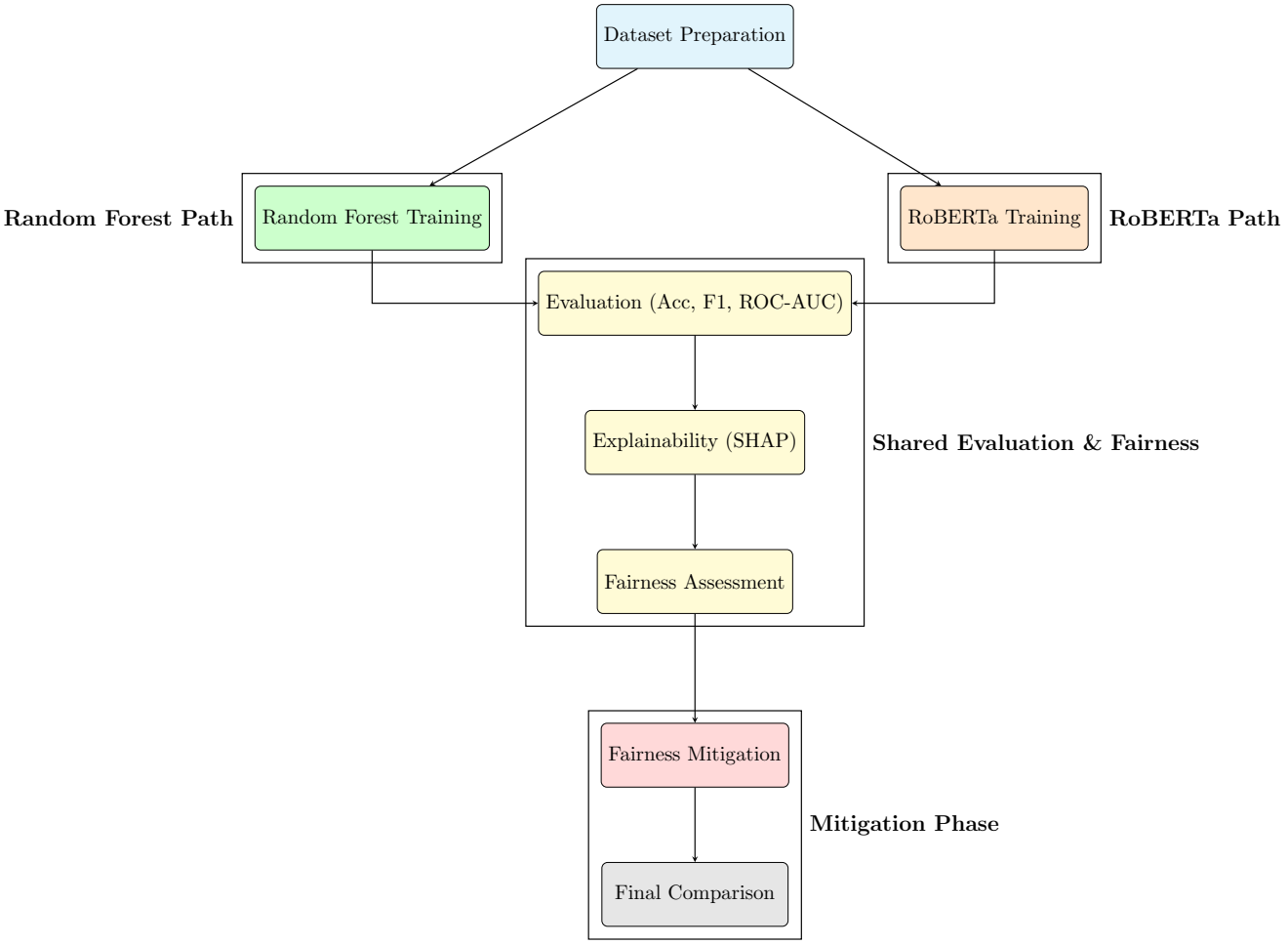


Figure 2: Methodology pipeline comparing Random Forest and RoBERTa models for AD detection, including evaluation, SHAP explainability, fairness assessment, and mitigation.

3.1 Datasets

This thesis uses two datasets that are available to researchers through DementiaBank upon request and approval: the Pitt Corpus (Becker et al., 1994) and the ADReSS Challenge dataset (Luz

et al., 2020). Access to these datasets is restricted and regulated by data use agreements to protect participant privacy and ethical research practices due to the sensitive nature of the recording and the associated metadata. Pitt Corpus is funded by NIA grants AG03705 and AG05133.
The Pitt Corpus was developed by the University of Pittsburgh as part of the DementiaBank project, a long-term research initiative supported by the National Institute of Aging. It is one of the most comprehensive collections of spontaneous speech recordings from this population and was created to facilitate the study of language and communication in people with dementia. The dataset consists of speech samples collected during the "Cookie Theft" picture description task (Figure 3), which is a component of the Boston Diagnostic Aphasia Examination (BDAE) (Fritsch et al., 2019).

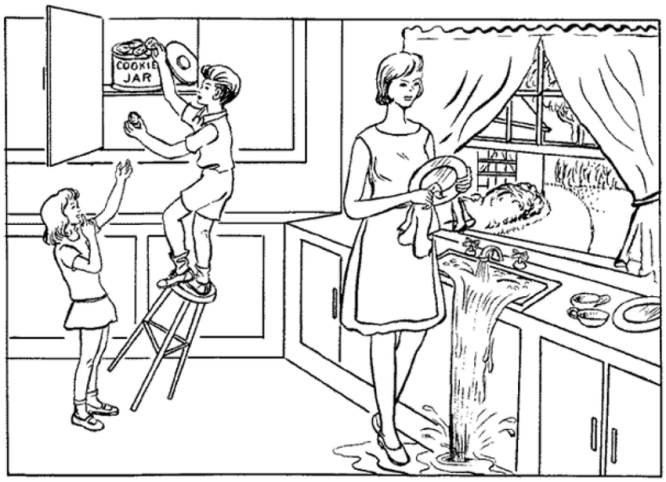


Figure 3: The "Cookie Theft" picture from the Boston Diagnostic Aphasia Examination (Goodglass & Kaplan, 1983)

The task is commonly used in neuropsychological assessments to obtain natural but structured language samples. The Pitt Corpus includes a heterogeneous group of participants representing a variety of cognitive conditions. These include probable AD, possible AD, MCI, vascular dementia, memory impairment and also healthy controls. Furthermore, the corpus contains longitudinal data for a large number of participants, some of whom provide multiple recordings over the course of twenty years. This thesis keeps every participant that is available from the Pitt Corpus "Cookie" directory, in contrast to earlier research that limits analysis to a filtered subset of participants, such as those who have been diagnosed with probable AD only (Shamei et al., 2023).

To ensure complete demographic coverage, missing age values were manually added using information from the official DementiaBank demographic spreadsheet associated with the Pitt Corpus. In cases where the age was missing or ambiguous in the raw transcript files (e.g., formatted as "69;00." or marked as an educated estimate), the entry age in the spreadsheet was used instead. This value reflects the participant's age at the time of the first recorded session, which directly corresponds to the analyzed speech sample. For the purpose of reporting dataset statistics, such as age and gender distributions in Table 2, only one transcript per participant was included to avoid duplication. In these cases, the transcript from the participant's first session was selected to represent them in the

overall demographic summary.

Table 2: Distribution of the participants from the Pitt Corpus by age group, gender, and diagnosis (AD: Alzheimer's Disease and HC: Healthy Control). Only the earliest transcript per participant is used, and bins reflect 5-year intervals from 45 to 90 years.

Age Interval	**AD**		**HC**	
	Male	Female	Male	Female
$[45, 50)$	0	1	0	3
$[50, 55)$	4	1	5	4
$[55, 60)$	6	8	6	13
$[60, 65)$	10	13	8	8
$[65, 70)$	12	24	10	15
$[70, 75)$	12	26	8	11
$[75, 80)$	19	29	3	4
$[80, 85)$	4	14	1	0
$[85, 90)$	0	10	0	0
Total	**67**	**126**	**41**	**58**

The ADReSS Challenge dataset is a subset of the Pitt Corpus. It was introduced as part of the 2020 Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) Challenge. It was designed to allow fair comparison of ML models for dementia detection by controlling for interfering variables like sound quality and gender and age distribution. The dataset contains a balanced training and testing sets as can be seen in Table 3.

Table 3: Distribution of the participants from the ADReSS dataset by age group, gender, and diagnosis (AD: Alzheimer's Disease and HC: Healthy Control). Each participant is represented by a single transcript, and age bins span 5-year intervals from 50 to 80 years.

Age Interval	**AD**		**HC**	
	Male	Female	Male	Female
$[50, 55)$	2	0	2	0
$[55, 60)$	7	6	7	6
$[60, 65)$	4	9	4	9
$[65, 70)$	9	14	9	14
$[70, 75)$	9	11	9	11
$[75, 80)$	4	3	4	3
Total	35	43	35	43

3.2 Preprocessing

The transcript files were originally annotated using the Codes for the Human Analysis of Transcripts (CHAT), a standardized system widely used within the Child Language Data Exchange System (CHILDES) database for analyzing conversational data. The CHILDES database is a repository of transcript, audio, and video data used in the study of first language acquisition. This coding scheme captures a range of linguistic features, allowing for consistent annotation across different studies. For this thesis, only the CHAT-annotated textual transcripts of participants' responses to the Cookie Theft picture task were used, providing the linguistic data for further analysis.

The raw transcript files were processed using a custom script to extract relevant metadata and clean the participant's speech. Some preprocessing steps were inspired by publicly available code shared in a previous study (Searle, 2020). Each file was analyzed to isolate the participant's utterances, identified by the `PAR:` level in the CHAT annotation format, while filtering out contributions from other speakers and unrelated metadata. For each utterance, disfluency markers such as repetitions `[//]`, retractions `[/]`, false starts `[/-]`, errors `[*]`, and abandoned phrases `[+...]` were counted prior to the text cleaning step. This allowed quantitative fluency indicators to be preserved.

In addition to the cleaned transcripts, key metadata was extracted from the header section of each CHAT file. This included the participant's gender and age available, as well as Mini-Mental State Examination (MMSE) scores for participants with AD when present. For each participant, counts of disfluencies, pauses, and utterances were recorded as additional features to support later linguistic analysis. When age information was missing or incorrectly formatted in the transcript headers, a manual correction step was added as explained in the previous section. The final cleaned data for both datasets (Pitt Corpus and ADReSS) was saved in tab-separated format for use in the next stages of feature extraction.

3.3 Linguistic Feature Engineering

A total of 32 linguistic features were extracted from each transcript to capture a wide range of lexical, syntactic, and semantic features in participants' speech. These features were selected to reflect fluency, lexical richness, syntactic complexity, and discourse content, drawing on previous research on language decline in dementia as well as established NLP techniques. In particular, the feature set was guided by the methodology proposed by (Fraser et al., 2016) and expanded based on findings by Fraser and Hirst (2016) and Eyigoz et al. (2020).

A custom Python script was developed to automate feature extraction, by combining standard NLP libraries with task-specific logic tailored to the Cookie Theft picture description task. Each transcript was processed through a pipeline consisting of tokenization, lemmatization, PoS tagging, dependency parsing, and feature computation using the `spaCy` English language model for tokenization (Honnibal et al., 2020). Common stop words were removed using the NLTK corpus.

Several PoS-derived features were included, such as, noun, verb, auxiliary, adverb, pronoun and preposition ratios, as well as first- and third-person pronoun usage. These features have been shown

to reflect referential specificity and syntactic simplification in AD (Wen et al., 2023). Although in this thesis sequential PoS n-grams were not added, higher-order syntactic structures were captured through CFG rule counts and parse tree depth.

Lexical richness was assessed using metrics such as type-token ratio (TTR), Brunet's index, Honoré's statistic, and Moving Average TTR (MATTR). These measures help capture vocabulary diversity and repetition frequency, with lower values indicating reduced lexical diversity, which is a known marker of AD speech (Fraser et al., 2016).

One semantic metric was the Information Unit (IU) count, which measures task-relevant content by counting expected keywords (e.g., "girl", "cookie jar", "kitchen") specific to the Cookie Theft description task. This metric is inspired by the Correct Information Unit (CIU) framework proposed by Nicholas and Brookshire (1993), but adapted here to focus on content relevance rather than full functional accuracy or discourse coherence.
The utility of this metric is also supported by the cognitive-linguistic framework proposed by Cummings (2019), which identifies specific linguistic markers affected in AD, such as the omission of key visual details, reduced lexical specificity, and lack of mental state language. Therefore, IU Count reflects how effectively a participant refers to semantically and pragmatically salient elements in the picture. While not equivalent to a full discourse analysis, this approximation provides a computationally efficient estimate of semantic informativeness. A summary of the extracted features is shown in Table 4, grouped by functional category.

Table 4: Overview of linguistic features extracted from each transcript, grouped by category. These 32 features quantify fluency, syntactic structure, lexical richness, and semantic content relevant to cognitive function.

Category	Feature	Code Name	Description
Lexical Richness	Type-Token Ratio (TTR)	`type_token_ratio`	Proportion of unique words to the total words.
	Moving-Average TTR (MATTR)	`mattr`	Average TTR over sliding windows; reduces sensitivity to length.
	Brunet's Index	`brunet_index`	Lexical richness measure; lower values indicate higher richness.
	Honoré's Statistic	`honore_statistic`	Emphasizes rare word usage; higher = greater lexical richness.
	Unique Word Count	`unique_words`	Number of different words used.
Fluency and Disfluency	Disfluency Count	`disfluency_count`	Total disfluencies (e.g., "um", "uh") detected.
	Pause Count	`pause_count`	Number of pauses identified in the transcript.
	Repetition Score	`repetition_score`	Mean similarity between sentence vectors (cosine similarity).
	Utterance Count	`utterance_count`	Total number of utterances.
Syntactic Complexity	Mean Sentence Length	`mean_sentence_length`	Average number of words per sentence.
	Parse Tree Depth	`parse_tree_depth`	Max depth of syntactic parse trees.
	Clauses per Sentence	`clauses_per_sentence`	Average number of subordinate clauses per sentence.
	CFG Rule Counts	`cfg_np_to_nn`, `cfg_np_to_det_noun`	Frequency of selected grammar rules used.
Lexical Usage	Noun/Verb Ratio	`noun_verb_ratio`	Ratio of nouns to verbs; higher = object-centric language.
	Pronoun Ratio	`pronoun_ratio`	Ratio of all pronouns to total word count.
	Stopword Ratio	`stopword_ratio`	Fraction of stopwords in total word count.
	First-Person Pronoun Ratio	`first_person_ratio`	Ratio of first-person pronouns (e.g., "I", "we").
	Third-Person Pronoun Ratio	`third_person_ratio`	Ratio of third-person pronouns (e.g., "he", "they").
	Named Entity Count	`named_entity_count`	Number of named entities (e.g., names, places).
	Auxiliary Ratio	`auxiliary_ratio`	Fraction of auxiliary verbs (e.g., "is", "have").
	Preposition Ratio	`prep_ratio`	Fraction of prepositions in the text.
	Adverb Ratio	`adv_ratio`	Fraction of adverbs in the text.
Semantic Content	Idea Density	`idea_density`	Ratio of content-bearing POS (noun, verb, adj, adv) to total words.
	Content Word Density	`content_density`	Proportion of informative POS to total words.
	Information Unit (IU) Count and IU Density	`IU_count`, `IU_density`	Count and normalized frequency of IU-related words (e.g., "girl", "cookie").
Redundancy and Repetition	Compression Ratio	`compression_ratio`	Ratio of compressed to original size; lower = more redundancy.
	Repetition Score	`repetition_score`	Mean sentence similarity; higher = more repetition.

Syntactic complexity was measured using features such as average sentence length, parse tree depth, clause density (i.e., clauses per sentence), and the number of context-free grammar (CFG)

production rules (e.g., `NOUN → DET NOUN`) as established by Fraser et al. (2016). Together, these features help understand the complexity of the participants' speech. To assess fluency, the pipeline counted several disfluency markers based on the CHAT transcription system, as previously stated. Pause markers, which were classified by their duration and position, were also included to estimate the level of hesitation.

Semantic and discourse-level content features included idea density (i.e., the ratio of content words to total words), the number of IU related keywords, pronoun usage ratios (e.g., first vs third person pronouns), and named entity frequency. Additionally, a compression ratio was also added using `zlib` to estimate information redundancy, with lower values indicating more repetitive or compressed speech.

Finally, discourse-level repetition was measured by computing pairwise cosine similarities between sentence vectors as explained in Fraser and Hirst (2016). This provides an estimate of how often ideas or structures are reused throughout a transcript. All features were computed from the cleaned and lemmatized transcripts, and the resulting values were stored with the participants' metadata for use in downstream classification models.

3.3.1 Statistical Comparison of Linguistic Features

To identify significant differences in linguistic features between AD and Control patients, a Welch's two-sample t-test was performed for the ADReSS and Pitt datasets. Welch's t-test was chosen over the standard Student's t-test because it does not assume equal population variances or sample sizes. Each feature was tested independently, and the results were ranked by p-value. Features like IU_density, pronoun_ratio, and adv_ratio showed statistically significant differences ($p < 0.05$) for both datasets, suggesting that these features are consistent linguistic markers of AD.

The results, shown in Table 5, indicate that several features differ significantly between the groups in both datasets. Notably, adv_ratio, auxiliary_ratio, IU_count, and pronoun_ratio showed p-values < 0.001 for both ADReSS and Pitt. These features align with established literature findings regarding reduced syntactic complexity and content informativeness in AD speech.

Table 5: Top linguistic features showing significant group differences in both ADReSS and Pitt datasets (Welch's t-test, $p < 0.05$). Control and AD group means are shown for interpretability.

Feature	ADReSS Control Mean	ADReSS AD Mean	Pitt Control Mean	Pitt AD Mean	ADReSS p-value	Pitt p-value
adv_ratio	0.0251	0.0463	0.0238	0.0379	<0.0001	<0.0001
IU_count	12.76	8.91	12.36	9.88	<0.0001	<0.0001
auxiliary_ratio	0.0855	0.0650	0.0831	0.0657	0.0015	<0.0001
pronoun_ratio	0.1143	0.1423	0.1126	0.1330	0.0027	<0.0001
prep_ratio	0.0902	0.0747	0.0924	0.0859	0.0061	0.0111
honore_statistic	1670.1	1408.4	1588.2	1449.5	0.0078	0.0002
IU_density	0.1197	0.0979	0.1176	0.1024	0.0124	<0.0001
pause_count	0.8889	1.5370	1.0782	1.5752	0.0459	0.0004
noun_ratio	0.2152	0.1927	0.2161	0.2007	0.0484	0.0001

Additionally, to verify the accuracy of the automated feature extraction process, a manual validation was performed on a subset of the data. Specifically, five transcripts from each group were randomly

selected and four features (IU_count, pronoun_ratio, adv_ratio, and pause_count) were manually counted and compared with the pipeline's output. This comparison confirms agreement with the automated extraction.

3.4 Model Setup

This thesis uses two models for AD detection: a Random Forest classifier and a RoBERTa-based model. Each model was trained separately on both datasets: ADReSS and Pitt.

3.4.1 Random Forest (RF)

This section describes the implementation of the Random Forest (RF) model across both datasets. In both cases, models were trained using manually engineered linguistic features derived from the transcribed speech samples as explained in Section 3.3.
For the ADReSS dataset, which includes predetermined training and testing splits, the initial RF model was trained using the complete training set (n=108 samples) and evaluated on the held-out test set (n=48 samples). Both sets are balanced, with equal number of AD and Control samples (54 each in training and 24 each in test). On average, participants in the training set produced 110.6 ± 63.2 words per transcript and in the test set 118.5 ± 81.7.

The initial model was trained using all available features. The feature importance scores obtained from this model were then used to rank and choose the top 18 features. The number of selected features (18) was determined empirically by experimenting with different values of k and observing the resulting model performance. Across both datasets, 18 features consistently resulted in the best accuracy. Following that, a final RF model was retrained using only these selected features and the evaluation was performed on the test set. The final model was intended for use in SHAP explainability analysis (refer to Section 4.2).

In contrast, the Pitt dataset contains a total of 549 samples with 243 Control and 306 AD cases, with a moderately imbalanced label distribution. On average, transcripts in this dataset contain 111.2 ± 60.5 words, with Control participants producing slightly longer samples (115.1 ± 60.9) than AD participants (108.1 ± 60). Since no predefined train/test split is provided, a 5-fold stratified cross validation approach was used to evaluate the model performance. In each fold 80% of the data (439 samples) was used for training and 20% (110 samples) for validation. An initial RF model was trained on the full dataset to rank feature importance. The top 18 features were then selected for training and evaluation within each fold. For each fold, a new RF model was developed and evaluated using the same specifications as ADReSS. Predictions and their probabilities were stored for later fairness analysis. Following cross-validation, a final model was built on the complete dataset using the reduced feature set for SHAP explainability.

During hyperparameter tuning, different number of trees (`n_estimators`) were evaluated: 100, 200, 300, 400, and 500. The best results were obtained with 100 trees, probably because of the small size of both datasets. Consequently all models were built using `RandomForestClassifier` from `scikit-learn` (Pedregosa et al., 2011), with `n_estimators=100`, `random_state=42`, and

`class_weight="balanced"` to handle class imbalance in Pitt Corpus. To improve interpretability, several visualizations were generated, including feature importance plots, correlation heatmaps, different SHAP plots, and individual decision trees.

3.4.2 RoBERTa

This section describes the implementation of RoBERTa models for both datasets. Unlike RF, which uses pre-engineered linguistic features, RoBERTa directly processes raw text, making use of contextualized word representations learned from large scale pretraining. Specifically, the `roberta-base` variant from Hugging Face's `transformers` library was used (Wolf et al., 2020).

RoBERTa is a transformer-based model that uses a multi-head self-attention mechanism and positional embeddings to encode input sequences. For text classification, which is the task used in this thesis, the final hidden state of the start-of-sequence token (`<s>`) is passed through a linear classification head to predict class probabilities. The model is fine-tuned end-to-end, so both pre-trained encoder and classification layer adapt to the AD vs Control classification task during training.

Although both `roberta-base` and `roberta-large` models were tested, the base model was the one ultimately selected due to several practical advantages. While `roberta-large` consists of 24 transformer layers and offers deeper contextual understanding, it is significantly more computationally expensive to fine-tune. Given the relatively small size of the ADReSS and Pitt datasets, `roberta-base`, which contains 12 layers, was found to achieve comparable performance with reduced training time and less risk of overfitting. Moreover, a learning rate sensitivity analysis was conducted using values of 1e-5, 2e-5, 3e-5. Given that 2e-5 and 3e-5 achieved similar results, the smaller value of 2e-5 was chosen aligning with common recommendations favoring lower learning rates for fine-tuning pre-trained models, as it also helps prevent overfitting.

For the ADReSS dataset which, as previously mentioned, contains predefined training and testing splits (108 training and 48 testing), RoBERTa was fine-tuned on the training set for five epochs using cross-entropy loss function and AdamW optimizer. Input sequences were tokenized with a maximum length of 256 tokens, including padding and truncation. An analysis of token lengths across the transcripts showed a minimum of 29 tokens and a maximum of 545, with a mean of 117.27. While a few samples exceeded 256 token limit and were therefore truncated, most of them were below that limit, given that 95% of the samples contained fewer than 230. Raising the maximum sequence length beyond 256 would have considerably increased computational expenses and training duration. After training, performance was evaluated on the test set using the standard metrics; accuracy, precision, recall, F1-score, and ROC-AUC. Additionally, prediction and probabilities were stored for downstream fairness and explainability analyses. Word-level SHAP values were computed using a Hugging Face pipeline and SHAP `Explainer` to interpret token contributions in individual predictions.

For the Pitt dataset (243 Control and 306 AD) a 5-fold stratified cross-validation setup was implemented, where 80% of the data (440 samples) was used for training and 20% (109 samples) for validation in each fold. In each fold, a new instance of the RoBERTa model was fine-tuned using the

training data and evaluated using the same standard metrics. Predictions and probabilities from all validation folds were aggregated to evaluate model performance and fairness across the full dataset. After cross-validation, a final RoBERTa model was trained on the full dataset for use in SHAP analysis and fairness mitigation. The tokenization and model settings, including the maximum sequence length of 256 tokens, as well as padding and truncation, were kept consistent with the ADReSS configuration. An analysis of token lengths in the Pitt dataset showed a similar distribution, with most transcripts containing fewer than 256 tokens (mean: 115.7 ± 61.3 and 95th percentile: 228).

Fairness analysis was conducted using AIF360 toolkit and with the same metrics applied to the RF models: SPD, EOD, AOD, and DI. ROC post-processing was applied to mitigate bias, and its effects on both fairness and accuracy were assessed. All RoBERTa models were implemented in PyTorch using Hugging Face Transformers, and training was conducted on GPU-enabled hardware when available.

3.5 Fairness approach

In order to identify and address potential biases in the classification models for detecting AD, this thesis uses the AIF360 Python toolkit developed by IBM (Bellamy et al., 2019). The focus is on achieving group fairness with respect to two demographic groups: gender (binary: male/female) and age (binary: individuals younger than 65 vs those aged 65 and older). These factors were chosen due to their significance in clinical settings (Alzheimer's Association, 2024).

Fairness was evaluated using standard metrics as defined in the AIF360 framework. These metrics include Statistical Parity Difference (SPD), which measures the disparity in the rate of favorable outcomes between privileged and unprivileged groups; Equal Opportunity Difference (EOD), which evaluates the discrepancy in true positive rates; Average Odds Difference (AOD), which considers the average differences in both true and false positive rates; and Disparate Impact (DI), which is the ratio of favorable outcome rates between groups. These metrics were calculated based on the model predictions, both before and after mitigation.

The methodology varied according to the characteristics of each dataset. The Pitt Corpus exhibits significant imbalances in group distribution across both age and gender, as can be seen in Table 2. As a consequence, a preprocessing bias mitigation strategy was applied. Specifically, a Reweighting algorithm was applied within a 5-fold stratified cross-validation pipeline. This algorithm adjusts the weights of training instances in order to reduce the influence of group imbalance between privileged and unprivileged groups. Additionally, the RF classifier was configured with `class_weight="balanced"` to address label imbalance between AD and control classes. In addition to preprocessing, a post-processing approach using Reject Option Classification (ROC) was used on the predictions obtained from the cross-validation. ROC modifies predictions in uncertain score regions near the decision boundary to improve fairness for unprivileged groups. ROC was applied with SPD as the target metric, with the decision boundary margin varying across different values to explore the trade-off between fairness and accuracy.

In contrast, the ADReSS dataset was constructed to minimize demographic imbalance across gender

and age as can be seen in Table 3. Therefore, no preprocessing mitigation was applied. All bias mitigation was performed post hoc on the test set predictions using ROC. Similarly to the approach used for Pitt, ROC was used with SPD as the target metric, and margin sweeps were conducted as well. All four fairness metrics (SPD, EOD, AOD, and DI) were tracked during the evaluation; however, mitigation depended exclusively on SPD. The findings from both pipelines are analyzed in Section 4.3.

4 Results

This section explores the findings from the experimental pipeline, which evaluates both Random Forest and RoBERTa on the ADReSS and Pitt datasets. The results are organized in three main sections: model performance, model explainability, and fairness. Each subsection addresses different aspects of the evaluation to explain how the models behave, which linguistic features influence their decisions, and how equitable their predictions are across demographic groups.

4.1 Model Performance

The classification performance of the two models, RF and RoBERTa, is evaluated on both datasets. Standard evaluation metrics including accuracy, precision, recall, F1-score, and ROC-AUC were used to assess how well the models can distinguish between control and AD cases. A summary of the results is presented in Table 6.

Table 6: Comparison of model performance on ADReSS and Pitt datasets. Bold values indicate the best performance for each metric.

Model	Dataset	Accuracy	Precision	Recall	F1-score	ROC-AUC
RF	ADReSS	**0.833**	**0.843**	**0.833**	0.832	0.889
RF	Pitt	0.720	0.718	0.709	0.710	0.800
RoBERTa	ADReSS	**0.833**	0.836	**0.833**	**0.833**	**0.901**
RoBERTa	Pitt	**0.794**	**0.804**	**0.802**	**0.794**	**0.876**

Across both datasets, RoBERTa outperforms the RF model on all metrics. On the ADReSS dataset, RoBERTa achieves an accuracy of 0.833 and an ROC-AUC of 0.901, matching RF's accuracy but surpassing its ROC-AUC of 0.889 by 1.2 points. This suggests that the transformer-based model captures better the language patterns directly from raw text compared to pre-engineered linguistic features.

In the Pitt dataset, the performance gap is especially notable in terms of ROC-AUC, where RoBERTa achieves 0.871 compared to 0.800 for RF, making it a 7.1 point difference. The accuracy difference here is larger than in ADReSS, with RoBERTa reaching 0.794 versus 0.720 for RF, a 7.4 point difference. While RoBERTa consistently shows higher average performance, RF exhibits slightly lower variance across most metrics, as shown in the cross-validation results Table 7.

Table 7: Cross-validation performance on the Pitt dataset (mean ± standard deviation). All metrics are macro-averaged.

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
RF	0.720 ± 0.031	0.718 ± 0.033	0.709 ± 0.032	0.710 ± 0.032	0.800 ± 0.019
RoBERTa	**0.794 ± 0.039**	**0.804 ± 0.041**	**0.802 ± 0.041**	**0.798 ± 0.036**	**0.876 ± 0.024**

4.2 Model Explainability

This section explores the interpretability of Random Forest and RoBERTa models for AD detection. For the Random Forest model, traditional interpretability techniques, such as feature importance, decision tree visualization, and correlation heatmaps, are used alongside SHAP to understand the impact of linguistic markers. For RoBERTa, model interpretation is primarily achieved through SHAP's text explainer, which highlights the tokens that contribute most to classification outcomes.

4.2.1 RF Explainability

One of the main reasons for choosing RF was its inherent interpretability, especially in contrast to more complex models like neural networks. To support this, several analysis were performed on the model on both datasets.

First, feature importance scores from the trained model were visualized to highlight which linguistic markers contributed the most to the predictions. As Figure 4 shows, both datasets have highly similar patterns of feature relevance. Auxiliary_ratio and honore_statistic are the highest contribution to the predictions across ADReSS and Pitt. Auxiliary_ratio reflect the use of auxiliary verbs in the transcript relative to all tokens and is often associated with grammatical complexity, which is more commonly observed in speakers without cognitive impairment.
Honoré's statistic, on the other hand, measures vocabulary complexity by giving more weight to words that appear only once. A higher value reflects a more diverse and complex vocabulary, which is usually reduced in individuals with AD. Both of these features are strongly supported in the literature as linguistic markers of cognitive decline (Williams et al., 2021). While both datasets agree on most feature importance rankings, there are some notable differences. For instance, content_density plays a more significant role in Pitt, while noun_ratio is more significant in ADReSS. These differences likely reflect variations in the dataset composition. Pitt includes a larger, more diverse group of participants, resulting in a greater variability in narrative style and structure.

Features such as noun_verb_ratio and mattr rank lower in both models, contributing less than other syntactic or lexical metrics. Moreover, not all features are present in both datasets. For instance, mean_sentence_length was selected among the top features only for ADReSS, while disfluency_count appears only in Pitt. Therefore, while some features appear less critical or are even absent from a given top-k ranking, they still contribute to the model's decision-making process. Figure 4 only shows the top 18 features out of the 32 used in total.

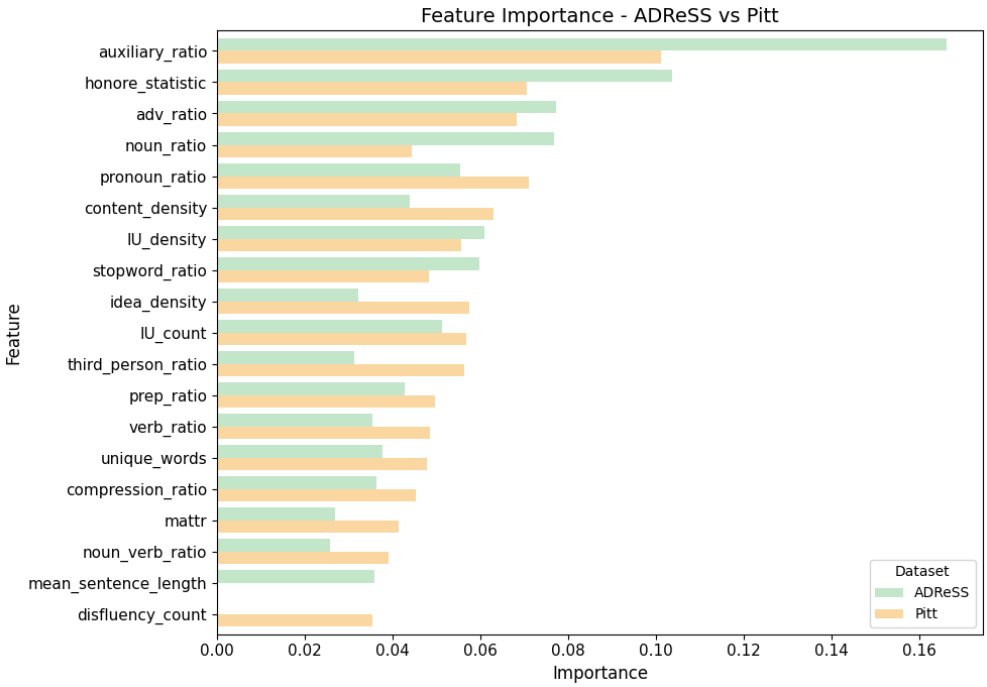


Figure 4: Comparison of ranked feature importance scores for the ADReSS and Pitt datasets using the Random Forest model. Features are ranked by their contribution to impurity reduction. The 18 most influential features (out of 32 in total), such as auxiliary_ratio and honore_statistic, appear consistently across datasets.

In addition to ranking features by importance, one of the decision trees trained within the RF model was visualized to better understand the decision-making structure (Figure 5). The tree provides a step by step illustration of how linguistic features influence the classification of individual samples. The root splits on content_density at a threshold of 0.373, indicating that utterances with lower lexical content density are more likely to be associated with AD. On the left branch (with lower content_density), the tree also evaluates pronoun_ratio and noun_verb_ratio. Increased use of pronouns can signal vague references which is commonly observed in AD speech, and a disproportionate number of nouns compared to verbs shows reduced sentence complexity. On the right branch (with high content_density), the model considers stopword_ratio, adv_ratio, and word_count. Shorter sentences may reflect reduced syntactic complexity which is a hallmark of language degradation in AD. Each split applied a learned threshold to maximize class purity. While this tree is only one component of the ensemble, it offers valuable insight into how linguistic features contribute to AD predictions.

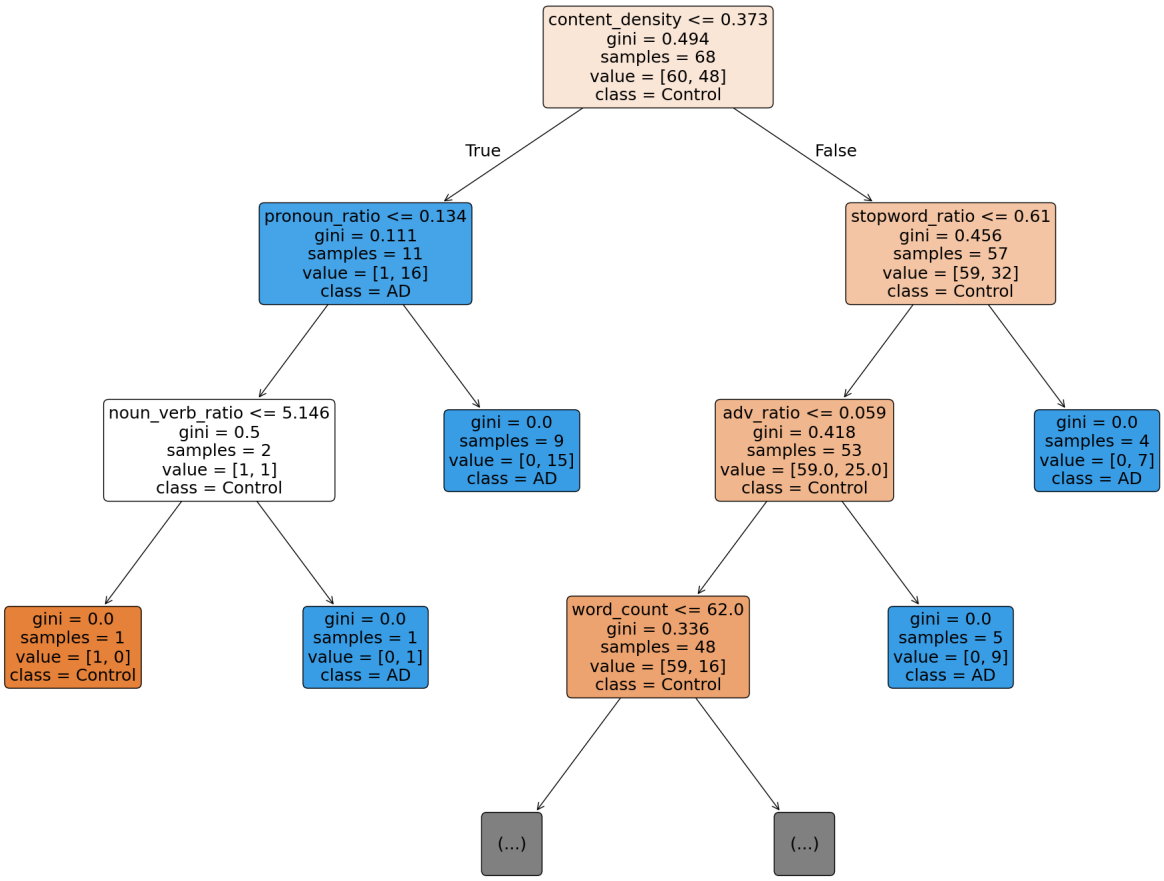


Figure 5: Visualization of one decision tree with depth 3 from the trained Random Forest model on the ADReSS dataset. The tree shows key linguistic splits, such as content_density and verb_ratio, are used to classify AD vs control speech samples.

Furthermore, to evaluate the relationships between the top 18 selected features, a correlation heatmap was generated (Figure 6). A strong negative correlation was observed between word_count and type_token_ratio with a value of -0.83. This indicates that larger texts tend to show less vocabulary diversity. Conversely, features like noun_ratio and IU_density have a strong positive correlation of 0.71, showing that individuals who include more task-specific words in their descriptions of the picture tend to use more nouns. The heatmap also shows potential redundancy between some features, such as unique_words and word_count (0.94) or compression_ratio and type_token_ratio (0.90), which may reflect overlapping lexical content, as both pairs capture related aspects of text length, complexity and lexical diversity.

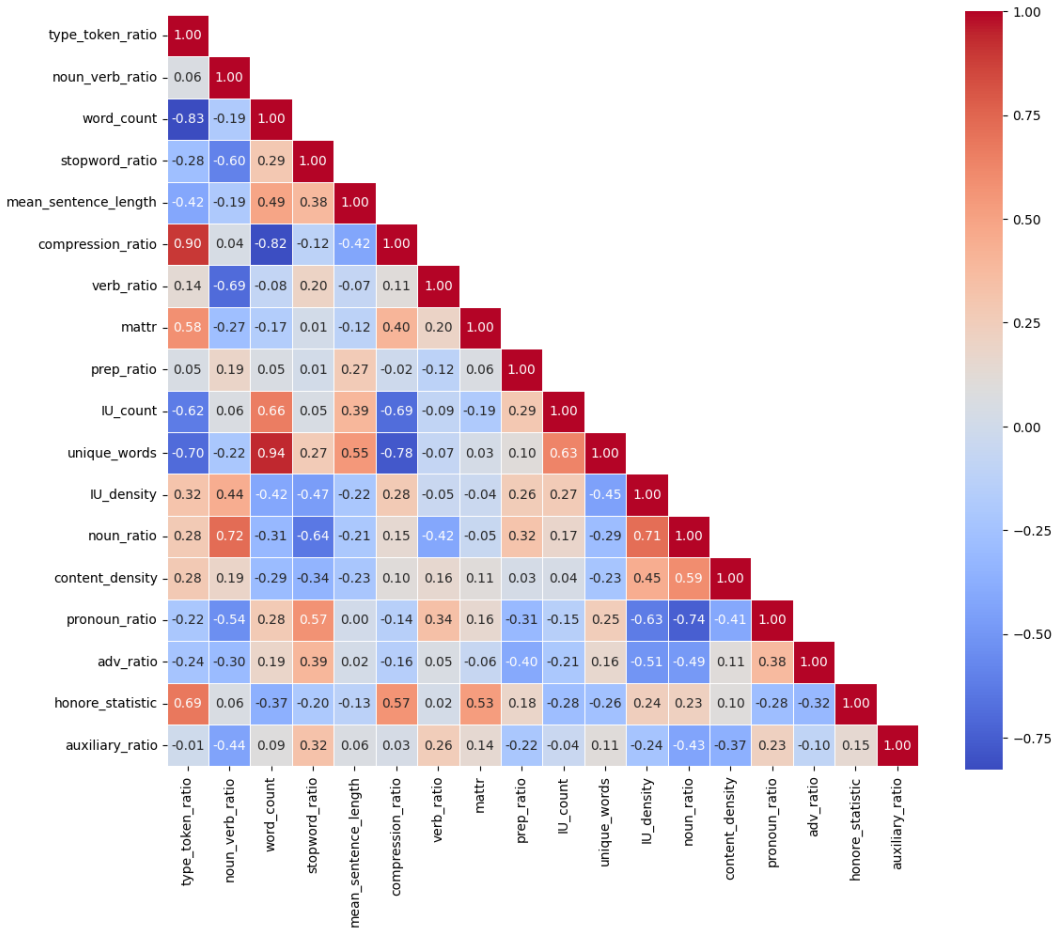


Figure 6: Pearson correlation heatmap of the top 18 features used in the RF model for the ADReSS dataset. Each cell represents the correlation coefficient between a pair of features; red indicating a strong positive correlation and blue, strong negative correlation. Notable relationships include high positive correlations between IU_density and noun_ratio (0.71) and strong negative between unique_words and compression_ratio (-0.78).

To complement traditional feature importance, SHAP values were used to provide a more detailed understanding of feature influence. The SHAP bar plot (Figure 13 in Appendix A 5.2) shows the overall impact of each feature. Linguistic features were grouped into six main categories: lexical, semantic, syntactic, pragmatic, CFG, and disfluency. This categorization aligns with standard linguistic theory as outlined in Jurafsky and Martin (2008). It can be seen that semantic and syntactical features have the most substantial influence on the model output. They include features like noun_ratio, IU_count, mean_sentence_length, among others. As previously mentioned, this is consistent with established linguistic markers of AD.

The beeswarm plot (Figure 14b in Appendix A 5.2) provides more detail by highlighting variation at the individual level. Features are ordered by importance (e.g., the impact of the feature on the model prediction). Positive values push the prediction toward AD and negative ones indicate a stronger association with control speech. For instance, higher auxiliary_ratio values tend to have negative SHAP values, suggesting they are associated with Control predictions. Conversely, lower

honore_statistic values (blue dots on the right side of the plot) tend to increase AD predictions, which is consistent with the reduced lexical diversity associated with AD patients.

A SHAP waterfall plot (Figure 7) was also generated to interpret an individual prediction made by the model. The transcription corresponding to this prediction is as follows:

"mhm . hm well sh shes um spillin the water from from washin uh her dishes . its its runnin over rather . in the the youngsters are are uh getting the jam . and in the meantime hes tiltin his chair. hm he is hes hes trying to get the cake down where sh sh I suppose she can share with him . and then and the mother is uh lis the water is runnin over laughs which she doesnt seem to be aware of it too much . and his chair is slippin out from uh the stool is slippin out from under him.".

The plot begins with the model's expected value (i.e., the mean predicted probability across all test samples), which for this case, is 0.436. SHAP values are then sequentially added or subtracted based on each feature's contribution. This process ends when it achieves a final predicted probability of 0.73 of this specific individual having AD. Noticeably, key features contributing positively to the AD prediction include honore_statistic, IU_density, and content_density. Auxiliary_ratio, on the contrary, pushes the prediction toward control, which aligns with existing clinical literature .

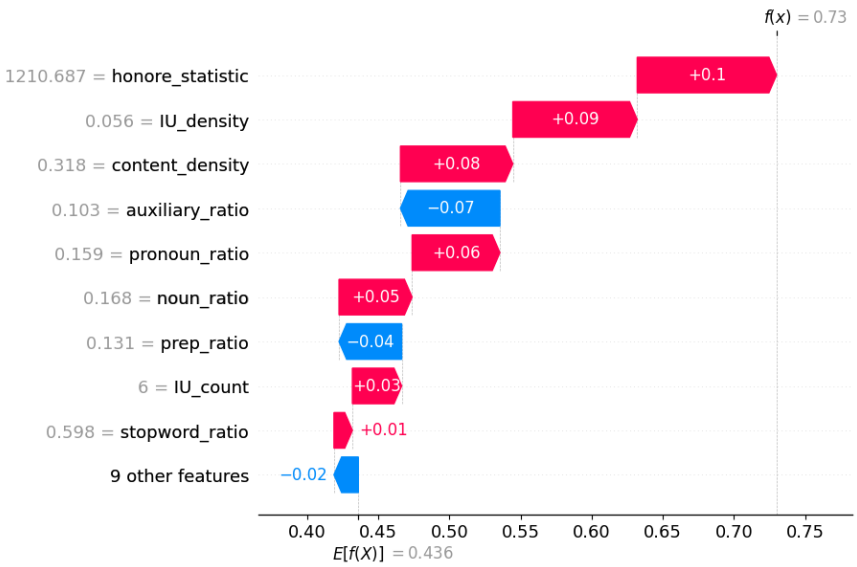
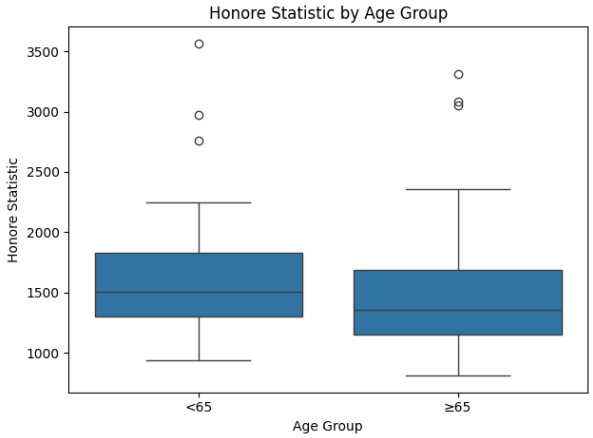


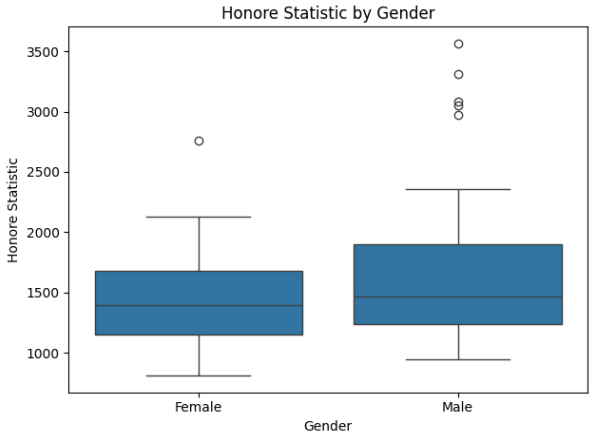
Figure 7: SHAP waterfall plot for an individual prediction from the RF model trained on the ADReSS dataset. The model starts from a base value of 0.447 and predicts a probability of 0.76 for AD. Each bar represents a feature's SHAP contribution to the final prediction. Features like honore_statistic strongly increase the AD prediction (in red), while auxiliary_ratio decreases it (in blue).

To investigate the role of honore_statistic further, an ablation study was conducted. Although the positive SHAP value of the honore_statistic initially seemed counterintuitive given that lexical richness is typically reduced in AD speech, the individual had a relatively low value (1210.687)

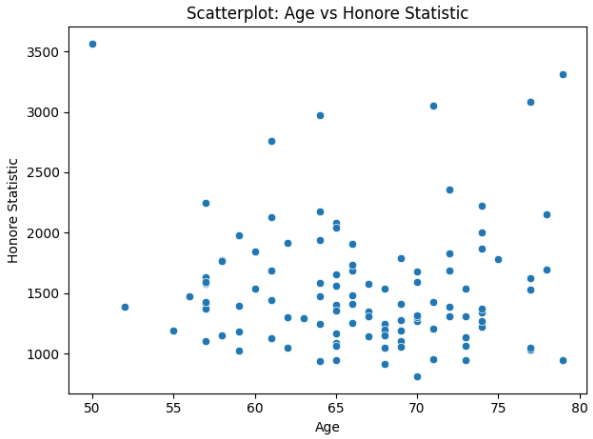
compared to the distribution of the AD group. In this study, the feature was removed and the model was retrained. This led to a drop in the model's overall accuracy, which strongly suggests that the feature is useful for helping the model make predictions. Finally, to assess potential fairness concerns, the relationship between honore_statistic and demographic variables was studied. Scatter plots and box plots (Figure 8) revealed no significant correlation with age or age group (two-sample t-test: $t = 1.431, p = 0.1570$). The scatterplot between age and honore_statistic also confirmed the lack of a linear relationship (Pearson correlation $r = -0.044$). However, when looking at gender, some differences were observed in the distribution of honore_statistic. A t-test confirmed a statistically significant difference between male and female participants ($t = -2.342, p = 0.0220$), indicating that honore_statistic tends to be lower in females.



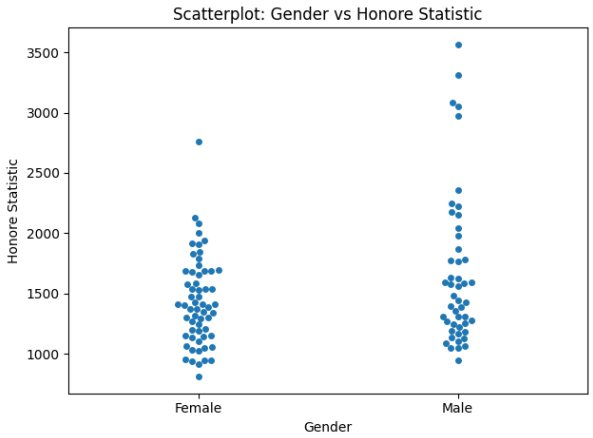
(a) Boxplot by age group.



(b) Boxplot by gender.



(c) Scatterplot: Age vs. Honoré statistic.



(d) Scatterplot: Gender vs. Honoré statistic.

Figure 8: Demographic analysis of the Honoré statistic. The boxplots show no notable differences across age groups ($t = 1.431, p = 0.1570$), while a moderate difference is observed between genders, with males showing higher values in average ($t = -2.342, p = 0.0220$). The scatterplot between age and honoré statistic confirms that, as no linear relationship between them exists (Pearson $r = -0.044$).

4.2.2 RoBERTa Explainability

Two complementary approaches were used to improve the interpretability of RoBERTa's predictions: SHAP was used for local interpretability, and BertViz was used to visualize the internal attention mechanism.

SHAP was applied using a Hugging Face text classification pipeline wrapped in a SHAP Explainer. This approach computes token-level attribution scores, showing how each word in a transcript contributes to the final prediction. Tokens are color-coded: red indicates a positive contribution to the predicted class and blue indicates a negative contribution away from the predicted class. For context, the final prediction score is shown alongside the base value, which represents the model's expected output before processing the input.

Figure 9 presents SHAP visualizations for one ADReSS test transcript.The model predicts the class as Control (label_0) with a probability of 0.97. Phrases such as "a saucer on the sink" and "the girl is touching her lips" contribute positively to the predicted class, while phrases like "theres a cup" or "her left shes" contribute negatively, pushing the prediction toward AD. The tool allows to toggle between label classes to observe SHAP values for the other classification.

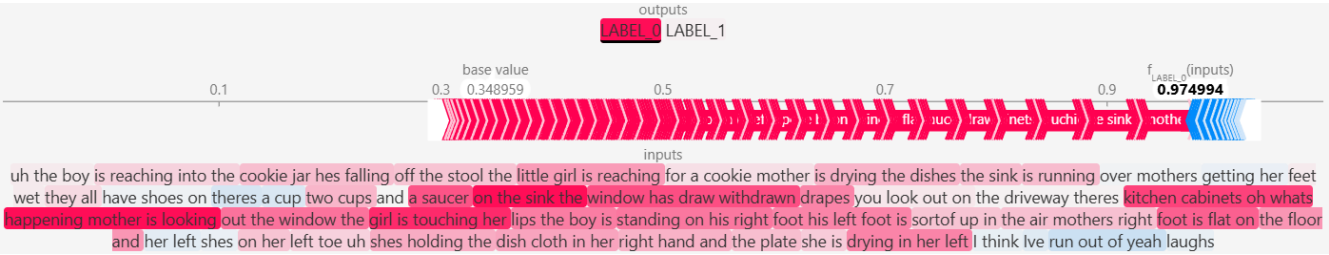


Figure 9: SHAP explanation for a prediction made on the ADReSS transcript. Red words contribute positively to the predicted class (Control) while blue words push it toward AD. The overall output probability is 0.97 toward label_0.

A second SHAP example in Figure 10 shows a transcript from the Pitt dataset. Again, the model predicts control with a high confidence (0.975). Phrases such as "reaching for a cookie" or "the window is open" push the prediction positively toward the control class. On the other hand, phrases like: "hes falling" push it more toward AD.

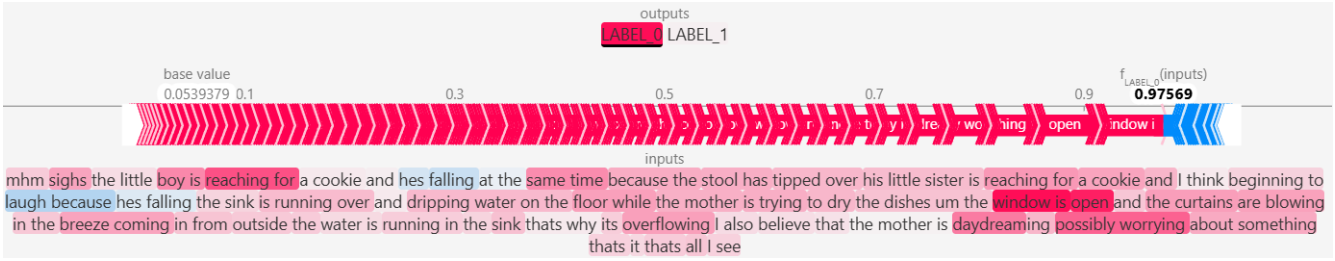


Figure 10: SHAP visualization for a Pitt sample predicted as Control. Key tokens such as "reaching for a cookie" support the decision. Tokens in red push the prediction toward Control, while blue tokens push the prediction toward AD.

Beyond token-level attribution, BertViz was used to examine RoBERTa's attention mechanisms. RoBERTa processes input sequences by computing attention weight between different tokens across multiple layers. These weights indicate the importance the model gives to one word when processing another. BertViz shows how different words are related through attention.

Figure 11 shows an example of the attention mechanism at Layer 0 for the tokens "boy" and "spilling". The lines radiating from these words to other words show the attention scores. Thicker, more opaque lines indicate higher attention weights, meaning the model focuses more on connected words when processing "boy" or "spilling". For "boy" the model shows strong connections to words like "this" and "is", indicating that the model is most likely creating a phrase internally. For "spilling" the model shows strong connection to "waters" and "over". This makes sense, as "water spilling over" forms a coherent semantic unit. The tool allows to toggle between different layers to observe the behavior.

(a) Attention for token *"boy"*.

(b) Attention for token *"spilling"*.

Figure 11: Attention maps from RoBERTa's Layer 0 for two key tokens from a transcript in the ADReSS dataset. Edges represent attention strengths. Thicker lines indicate stronger attention between tokens.

Figure 12 shows a misclassified case; an AD patient whose transcript was incorrectly predicted as Control, with very high confidence (96.86%). The SHAP explanation indicates that the model's decision was significantly influenced by concrete phrases such as "a plate in her hands" or "plate on the counter". These features, while linguistically rich, are not necessarily uncharacteristic of early-speech AD and may have led the model to an incorrect prediction. The speech as a whole lacks a clear storyline and grammatical cohesion, which are features affected in AD patients.

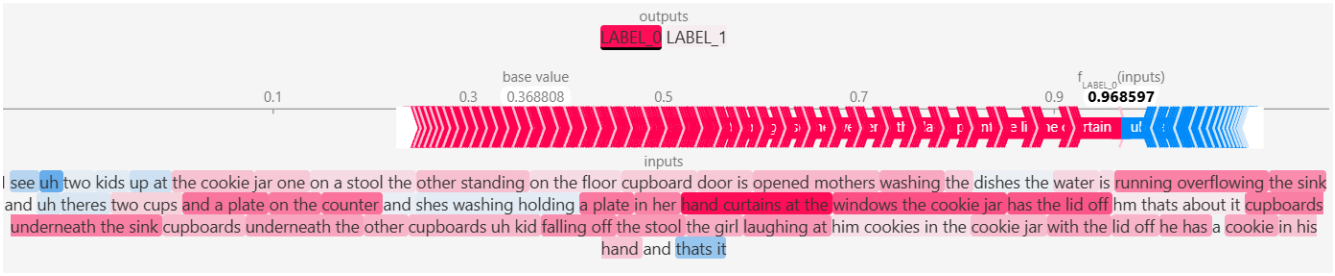


Figure 12: SHAP explanation for a misclassified AD transcript from the ADReSS dataset. The model incorrectly predicts Control with 96.86% confidence. Linguistically rich and descriptive phrases contribute significantly to the prediction, while disfluencies like "uh" are not sufficient to counteract it.

This example shows a limitation of the model's decision-making, as it does not capture broader discourse-level problems in the speech, such as the syntactic structure or the narrative coherence. Consequently, the model may seem to make reasonable decisions even when its underlying reasoning is inconsistent with clinical expectations. SHAP can be used to reveal that the model's focus does not necessarily align with clinically relevant factors.

4.3 Fairness

This section analyzes the fairness of the model predictions with respect to sensitive/protected attributes, such as gender and age group. Results from both bias detection and bias mitigation are presented. Standard fairness metrics are computed before and after applying mitigation techniques. These metrics include SPD, EOD, AOD, and DI. To compute these metrics using AIF360, one outcome label must be defined as "favorable". In this analysis, predictions labeled as AD (label = 1) are considered favorable. This reflects the clinical importance of minimizing missed diagnoses (i.e., maximizing recall), as undetected cases of AD may delay access to treatment. At the same time, this framing introduces a trade-off, since false positives can also have serious implications, such as unnecessary stress for patients or invasive and costly follow-up procedures. Therefore, AD is chosen as the favorable class for consistency within the fairness framework while acknowledging the balance between recall and specificity (true negative rate) in clinical decision-making. The goal of the following sections is to determine whether the models exhibit discriminatory behavior and how effectively it can be reduced without sacrificing accuracy.

4.3.1 Bias Detection

The fairness evaluation results for the RF and RoBERTa models before bias mitigation are presented in Table 8 and Table 9, respectively.

Table 8 shows the fairness metrics for the RF model. On the ADReSS dataset there are clear disparities for the age group. For gender, all three difference-based metrics, SPD, EOD, and AOD, are negative but relatively small. This indicates that male participants (the unprivileged group) are slightly underdiagnosed compared to female participants (the privileged group). The DI value of 0.967, further supports this small bias. In contrast, age-related disparities are more pronounced. The positive SPD, EOD, and AOD values suggest that younger participants (those under 65, the unprivileged group) are more likely to receive a positive AD diagnosis than the older group. The DI value of 1.667, points out a bias against the older group, the privileged one.

Table 8: Fairness metrics for Random Forest model before bias mitigation.

Metric	ADReSS		Pitt	
	Gender	Age	Gender	Age
Statistical Parity Difference	-0.014	0.222	-0.012	-0.084
Equal Opportunity Difference	-0.042	0.222	-0.112	0.031
Average Odds Difference	-0.014	0.222	-0.006	0.029
Disparate Impact	0.967	1.667	0.980	0.870

On the Pitt dataset, the RF model shows improved fairness, especially with regard to age. For gender, SPD and AOD values are close to zero, and the DI value of 0.980 suggests parity between

genders. However, EOD indicates a slight underdiagnosis of male participants. With regard to age, the results indicate a mild bias against the younger group. Although the SPD, EOD, and AOD values are nearly neutral, the DI metric suggests that participants under 65 are slightly less likely to receive an AD diagnosis. However, this disparity is much less pronounced than in the ADReSS dataset.

Turning to the RoBERTa model, Table 9 shows the corresponding fairness metrics.
On the ADReSS dataset, the gender-related metrics are close to zero. However, the negative EOD value of -0.119 and the DI value of 0.818, suggest minimal bias against male participants. Age-related disparities are reduced compared to the RF model. The SPD and AOD are both 0.067, indicating good fairness. Nevertheless, the DI value of 1.154 still indicates a slight bias against the older participants.

Table 9: Fairness metrics for RoBERTa model before bias mitigation.

Metric	**ADReSS**		**Pitt**	
	Gender	**Age**	**Gender**	**Age**
Statistical Parity Difference	-0.091	0.067	-0.029	-0.213
Equal Opportunity Difference	-0.119	-0.022	-0.203	-0.076
Average Odds Difference	-0.091	0.067	-0.016	-0.059
Disparate Impact	0.818	1.154	0.940	0.599

On the Pitt dataset, gender fairness shows relatively worse results for RoBERTa. The EOD value of -0.203 suggests a notable bias against male participants. Age-related disparities are more pronounced. All three different metrics are negative, and the DI drops to 0.599, indicating substantial underdiagnosis of the younger group.

4.3.2 Bias Mitigation

Given the disparities observed in the previous section, this section evaluates the impact of bias mitigation techniques. Two commonly used strategies were used: ROC and Reweighting. Fairness metrics were recalculated after mitigation to assess the effectiveness in reducing the bias and also the impact on the accuracy.

Table 10 presents the results after applying ROC and Reweighting on the RF model. For the ADReSS dataset, applying ROC substantially improved fairness metrics for both gender and age groups. In the gender group, all metrics moved closer to their ideal values: SPD and AOD were reduced to zero, and DI moved to 1. Although EOD remained slightly negative, it was significantly reduced. Notably, these improvements did not reduce accuracy, which remained at 0.833. For the age group, ROC led to substantial improvements across most metrics, significantly reducing the bias against the older group. However, EOD remained moderately high, indicating that some disparities

against the older group still remain. Despite this, the fairness gain came at the cost of a decrease in accuracy, dropping from 0.833 to 0.750.

Table 10: Accuracy before and after mitigation and fairness metrics after mitigation for the Random Forest model.

Dataset	**Subgroup**	**Accuracy** (Before | After)	**SPD**	**EOD**	**AOD**	**DI**
ADReSS (ROC)	Gender	0.833 | 0.833	0.000	-0.028	0.000	1.000
ADReSS (ROC)	Age	0.833 | 0.750	0.000	0.133	0.000	1.000
Pitt (Reweight)	Gender	0.720 | 0.727	0.017	-0.019	0.007	1.050
Pitt (Reweight)	Age	0.720 | 0.723	-0.079	0.000	0.034	0.884
Pitt (ROC)	Gender	0.720 | 0.727	0.009	-0.089	0.015	1.015
Pitt (ROC)	Age	0.720 | 0.720	-0.006	0.119	0.123	0.985

On the Pitt dataset for RF, Reweighting was moderately effective for the gender group. Fairness metrics improved overall, particularly EOD, and accuracy slightly improved from 0.720 to 0.727. This indicates a reduction in the bias against male participants. For the age group, Reweighting also improved all metrics, with SPD showing the most notable change. However, some bias against the younger participants remained, as the DI value improved but still remained below the ideal threshold of 1. Accuracy also improved from 0.720 to 0.723.

ROC mitigation on the RF model for the Pitt dataset had mixed results. For the gender group, it maintained relatively strong fairness metrics with only a small drop in accuracy (from 0.738 to 0.729). Most fairness values were near ideal, effectively removing the bias against the male participants. For the age group, ROC reduced the original slight bias against younger participants, as shown by improvements in SPD and DI values. However, this came at the cost of introducing a new bias against the older group, as seen in the reversal of the EOD and AOD values. Accuracy remained unchanged at 0.720.

Table 11 shows the results after applying ROC mitigation to the RoBERTa model. On the ADReSS dataset, ROC was generally effective. For the gender group, fairness improved substantially, with all metrics achieving near perfect fairness and accuracy increasing from 0.833 to 0.875. Similarly, for the age group, ROC improved most fairness metrics and notable reduced bias against younger participants. However, EOD slightly worsened. Notably, these fairness improvements did not impact accuracy, which remained at 0.833.

Table 11: Accuracy and fairness metrics before and after ROC mitigation for the RoBERTa model.

Dataset	**Subgroup**	**Accuracy** (Before | After)	**SPD**	**EOD**	**AOD**	**DI**
ADReSS (ROC)	Gender	0.833 | 0.875	-0.007	-0.028	-0.007	0.985
ADReSS (ROC)	Age	0.833 | 0.833	0.000	-0.089	0.000	1.000
Pitt (ROC)	Gender	0.794 | 0.800	0.003	-0.176	0.017	1.007
Pitt (ROC)	Age	0.794 | 0.778	-0.007	0.127	0.146	0.987

On the Pitt dataset, ROC mitigation improved fairness for the gender group, alongside a small increase in accuracy (0.794 to 0.800). Most metrics were close to ideal, though EOD remained biased against male participants with only marginal improvement. In contrast, ROC significantly reduced bias against the younger group, especially in SPD and DI. However, it introduced a new bias against the older participants, as seen in higher EOD and AOD values. Additionally, accuracy decreased slightly from 0.794 to 0.778.

5 Conclusion and Future Research

5.1 Conclusion and Discussion

This thesis investigated the detection of AD from transcribed speech using two types of models: a classical machine learning model (Random Forest) and transformer-based model (RoBERTa). The models were evaluated on two datasets, ADReSS and Pitt, to compare their performance, explainability, and fairness.

RoBERTa outperformed RF in terms of classification performance on both datasets, achieving higher accuracy and F1-scores due to its ability to use contextualized word embeddings from raw text. The performance gap was particularly notable on the Pitt dataset, where RoBERTa showed substantial improvements over RF, especially in terms of recall and ROC-AUC. On the ADReSS dataset, both models achieved the same accuracy and recall, however, RoBERTa achieved a higher ROC-AUC. Nevertheless, the RF classifier produced comparable results with a simpler architecture and offered advantages in interpretability. With further tuning, its performance could likely improve.

Explainability was a central focus of this thesis. RF provided both global and local interpretable results through feature importance rankings, SHAP-based visualizations, and decision tree visualizations. The model's explanations were relatively easy to link to known linguistic markers of AD. In contrast, RoBERTa's predictions were explained using token-level SHAP values and attention patterns using BertViz. While these tools highlight influential words, they are more difficult to interpret and lack the structured, feature-level insight RF offers. RoBERTa's explanations are usually scattered across tokens without clear aggregation, making them harder to relate to known linguistic markers of AD. Especially in clinical contexts, this makes RF's interpretability not only easier to work with, but also more trustworthy.

An ablation study was conducted to investigate the importance of the Honoré statistic. Although it initially showed some inconsistencies with literature, its removal reduced model performance, confirming its value for the model. Additionally, an exploratory demographic analysis showed that Honoré statistic values were lower for female participants. This is relevant because it suggests a potential interaction between the linguistic features and gender, which may contribute to model bias.

In response to the first research question, the results show that both gender and age influence the predictive performance of the models. This is clearly reflected in the significant disparities observed in the fairness metrics. On the ADReSS dataset, RF showed notable age-related bias, while gender-related bias was substantially less. RoBERTa showed lower overall bias on ADReSS, in particular with respect to age, suggesting that pretraining on large language data may help mitigate bias in small and balanced datasets, likely due to the limited variability of such data. However, this advantage did not hold for the Pitt dataset, which is larger and more imbalanced. For Pitt, RoBERTa, showed greater bias than RF for both gender and age, but more noticeable for age-related bias. Given the results, the fairness observed in ADReSS may be more reflective of dataset characteristics than the inherent fairness of the model. Across both datasets, the most consistent gender-related bias was observed against male participants. However, the direction of age-related bias varied by dataset: older participants were more disadvantaged in ADReSS, while

younger participants were more disadvantaged in Pitt.

Regarding the second research question, fairness-aware AI techniques, particularly Reject Option Classification (ROC) and Reweighting, were generally effective, though their impact varied. On the ADReSS dataset, ROC led to nearly perfect fairness outcomes for both models, with little or no loss in accuracy. In fact, RoBERTa even achieved improved accuracy for the gender subgroup. In contrast, the Pitt dataset, had mixed results. For RF, Reweighting moderately improved fairness for both gender and age groups, with slight gains in accuracy. ROC also helped, but in some cases it introduced new biases, in particular, a reversal of the age-related bias. For RoBERTa, ROC improved gender fairness but worsened EOD. Mitigating age bias resulted in a reduced accuracy and added new disparities against the older group. These results show that while fairness mitigation methods can significantly improve fairness, they may also shift bias across subgroups or affect accuracy.

In summary, this thesis shows that transformer-based models like RoBERTa achieve strong predictive performance in AD detection from speech, especially with larger datasets. However, classical models like RF remain strong candidates, offer advantages in interpretability, more stable fairness across datasets, and better transparency for clinical applications. Furthermore, the results highlight the importance of combining predictive performance with explainability and fairness analyses. These aspects are critical for developing ethical and trustworthy AI systems in healthcare.

5.2 Future Research

There are several ways to extend this work. First, the current models were trained and evaluated using datasets from a single task, the Cookie Theft picture description task. Future research could explore multitask learning frameworks that incorporate more diverse narrative tasks. Additionally, integrating acoustic features alongside linguistic features could enrich the input space and help models generalize better to different contexts.

Second, the datasets used in this thesis are relatively small for training deep learning applications. Transformer-based models like RoBERTa usually require large datasets to prevent overfitting. Data augmentation, such as data-to-text generation could be also explored to address this limitation.

Third, further investigation into features redundancy, as discussed in Section 4.2.1 could improve model generalization. For instance, correlations observed between different features such as clauses_per_sentence and mean_sentence_length or idea_density and content_density suggest overlapping. Redundant features may unnecessarily increase model complexity or even introduce noise. A more rigorous feature selection process by removing strongly overlapping features could lead to better results.

Fourth, the set of linguistic features could be expanded by using tools like CLAN from the TalkBank project. Using CLAN could lead to the inclusion of richer, more structured linguistic features, including morphological or prosodic cues, which could additionally improve the accuracy. Additionally, the optimization of the models in this thesis was limited because the primary focus was not on

achieving state-of-the-art performance. Future work could incorporate hyperparameter tuning (e.g., Grid Search or Bayesian Optimization) to achieve even better performance results.

Fifth, the fairness analysis revealed persistent biases, especially with regard to age. Although post-processing ROC mitigated some disparities, they introduced trade-offs in accuracy or shifted bias to the other group. Further research could explore alternative bias mitigation strategies, such as Adversarial Debiasing or Prejudice Remover, which are in-processing mitigation techniques. These techniques modify the model's internal learning process directly to reduce bias while preserving performance.

Sixth, this thesis focused only on binary classification to distinguish between AD and control participants. However, real-world diagnoses involve a spectrum of cognitive states, from healthy aging to MCI to AD. Even within the Pitt Corpus there are a range of diagnoses. Future work could incorporate multiclass classification to more accurately reflect clinical relevance.

Lastly, while this thesis used RoBERTa and RF to explore contrasts in model complexity and interpretability, future research could explore other architectures. This may include ensemble methods that combine classical and neural models or newer transformer-based models such as NeoBERT.

References

Alanazi, A. (2022). Using Machine Learning for Healthcare Challenges and Opportunities. *Informatics in Medicine Unlocked*, *30*, 100924. https://doi.org/10.1016/j.imu.2022.100924

Alzheimer's Association. (2024). 2024 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, *20*(5). https://www.alz.org/getmedia/76e51bb6-c003-4d84-8019-e0779d8c4e8d/alzheimers-facts-and-figures.pdf

Balagopalan, A., Eyre, B., Robin, J., Rudzicz, F., & Novikova, J. (2021). Comparing Pre-trained and Feature-Based Models for Prediction of Alzheimer's Disease Based on Speech. *Frontiers in Aging Neuroscience*, *13*, 635945. https://doi.org/10.3389/fnagi.2021.635945

Becker, J. T., Boller, F., Lopez, O. L., Saxton, J., & McGonigle, K. L. (1994). The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis [Funding support: NIA AG03705 and AG05133]. *Archives of Neurology*, *51*(6), 585–594. https://doi.org/10.1001/archneur.1994.00540180063015

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4:1–4:15. https://doi.org/10.1147/JRD.2019.2942287

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., & Walker, K. (2020). *Fairlearn: A toolkit for assessing and improving fairness in ai* (Technical Report No. MSR-TR-2020-32). Microsoft.

Chen, W., & Wang, H. (2013). Mild cognitive impairment: A concept useful for early detection and intervention of dementia. *Shanghai Archives of Psychiatry*, *25*(2), 119–120. https://doi.org/10.3969/j.issn.1002-0829.2013.02.009

Cummings, L. (2019). Describing the Cookie Theft picture: Sources of breakdown in Alzheimer's dementia. *Pragmatics and Society*, *10*, 151–174. https://doi.org/10.1075/ps.17011.cum

Das, D., Banerjee, D., Aditya, S., & Kulkarni, A. (2024). MATHSENSEI: A Tool-Augmented Large Language Model for Mathematical Reasoning. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 942–966. https://doi.org/10.18653/v1/2024.naacl-long.54

de Oliveira, J., Kucharska, E., Garcez, M. L., Rodrigues, M. S., Quevedo, J., Moreno-Gonzalez, I., & Budni, J. (2021). Inflammatory Cascade in Alzheimer's Disease Pathogenesis: A Review of Experimental Findings. *Cells*, *10*(10). https://doi.org/10.3390/cells10102581

Deng, H., Wang, C., Xin, L., Yuan, D., Zhan, J., Zhou, T., Ma, J., Gao, J., & Xu, R. (2024). WebCiteS: Attributed Query-Focused Summarization on Chinese Web Search Results with Citations. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15095–15114. https://doi.org/10.18653/v1/2024.acl-long.806

Eyigoz, E., Mathur, S., Santamaria, M., Cecchi, G., & Naylor, M. (2020). Linguistic markers predict onset of Alzheimer's disease. *EClinicalMedicine*, *28*, 100583. https://doi.org/10.1016/j.eclinm.2020.100583

Fraser, K. C., & Hirst, G. (2016). Detecting semantic changes in Alzheimer's disease with vector space models. *International Conference on Language Resources and Evaluation*. https://api.semanticscholar.org/CorpusID:17945535

Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2016). Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *Journal of Alzheimer's Disease*, *49*(2), 407–422. https://doi.org/10.3233/JAD-150520

Fritsch, J., Wankerl, S., & Noeth, E. (2019). Automatic Diagnosis of Alzheimer's Disease Using Neural Network Language Models, 5841–5845. https://doi.org/10.1109/ICASSP.2019.8682690

Fu, Y., Xu, L., Zhang, Y., Zhang, L., Zhang, P., Cao, L., & Jiang, T. (2024). Classification and diagnosis model for Alzheimer's disease based on multimodal data fusion. *Medicine*, *103*(52), e41016. https://doi.org/10.1097/md.0000000000041016

Goodglass, H., & Kaplan, E. (1983). *Boston Diagnostic Aphasia Examination Booklet*. Lea & Febiger.

Honnibal, M., Montani, I., Landeghem, S. V., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python [Version 2.0]. https://doi.org/10.5281/zenodo.1212303

Jurafsky, D., & Martin, J. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2nd). Prentice Hall.

Kavé, G., & Goral, M. (2016). Word retrieval in picture descriptions produced by individuals with alzheimer's disease. *Journal of Clinical and Experimental Neuropsychology*, *38*(9), 958–966. https://doi.org/10.1080/13803395.2016.1179266

Lindsay, H., Tröger, J., & König, A. (2021). Language Impairment in Alzheimer's Disease—Robust and Explainable Evidence for AD-Related Deterioration of Spontaneous Speech through Multilingual Machine Learning. *Frontiers in Aging Neuroscience*, *13*. https://doi.org/10.3389/fnagi.2021.642033

Liu, N., Luo, K., Yuan, Z., & Chen, Y. (2022). A Transfer Learning Method for Detecting Alzheimer's Disease Based on Speech and Natural Language Processing. *Frontiers in Public Health*, *10*. https://doi.org/10.3389/fpubh.2022.772592

Liu, N., Yuan, Z., & Tang, Q. (2022). Improving Alzheimer's Disease Detection for Speech Based on Feature Purification Network. *Frontiers in Public Health*, *9*. https://doi.org/10.3389/fpubh.2021.835960

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. https://arxiv.org/abs/1907.11692

Liu, Y.-L., Feng, R., Yuan, J.-H., & Ling, Z.-H. (2024). Clever Hans Effect Found in Automatic Detection of Alzheimer's Disease through Speech. *arXiv preprint arXiv:2406.07410*. https://doi.org/10.48550/arXiv.2406.07410

Luz, S., Haider, F., de la Fuente, S., Fromm, D., & MacWhinney, B. (2020). Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge. *Proceedings of INTERSPEECH 2020*, 2162–2166. https://doi.org/10.21437/Interspeech.2020-2571

Mao, C., Xu, J., Rasmussen, L., Li, Y., Adekkanattu, P., Pacheco, J., Bonakdarpour, B., Vassar, R., Shen, L., Jiang, G., Wang, F., Pathak, J., & Luo, Y. (2023). AD-BERT: Using Pre-Trained Language Model to Predict the Progression from Mild Cognitive Impairment to Alzheimer's

Disease. *Journal of Biomedical Informatics*, *144*, 104442. https://doi.org/10.1016/j.jbi.2023.104442

Mosteiro, P., Kuiper, J., Masthoff, J., Scheepers, F., & Spruit, M. (2022). Bias discovery in machine learning models for mental health. *Information*, *13*(5), 237. https://doi.org/10.3390/info13050237

Nicholas, L. E., & Brookshire, R. H. (1993). A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of Speech and Hearing Research*, *36*(2), 338–350. https://doi.org/10.1044/jshr.3602.338

Olmos-Villaseñor, R., Sepulveda-Silva, C., Julio-Ramos, T., Fuentes-Lopez, E., Toloza-Ramirez, D., Santibáñez, R. A., Copland, D. A., & Mendez-Orellana, C. (2023). Phonological and Semantic Fluency in Alzheimer's Disease: A Systematic Review and Meta-Analysis. *Journal of Alzheimer's Disease*, *95*(1), 1–12. https://doi.org/10.3233/JAD-221272

Ortiz, K. Z., De Lira, J. O., Minett, T. S., & Bertolucci, P. H. (2024). Language impairments in Alzheimerś disease: What changes can be found between mild and moderate stages of the disease? *Clinics (Sao Paulo)*, *79*, 100412. https://doi.org/10.1016/j.clinsp.2024.100412

Pandey, H. (2022). Comparison of the usage of fairness toolkits amongst practitioners: Aif360 and fairlearn [Bachelor Thesis]. http://resolver.tudelft.nl/uuid:4ef11035-2f60-436f-85f9-7b9bed73b66d

Park, S.-H., Kwon, K. J., Kim, M. Y., Kim, J.-H., Moon, W.-J., Ryu, H. J., Jang, J. W., Moon, Y., & K-ARPI. (2023). Diagnostic tools for alzheimer's disease: A narrative review based on our own research experience [Epub 2023 Feb 14]. *Dementia and Neurocognitive Disorders*, *22*(1), 16–27. https://doi.org/10.12779/dnd.2023.22.1.16

Parsapoor, M., Alam, M., & Mihailidis, A. (2023). Performance of machine learning algorithms for dementia assessment: impacts of language tasks, recording media, and modalities. *BMC Medical Informatics and Decision Making*, *23*, 45. https://doi.org/10.1186/s12911-023-02122-6

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

Pessach, D., & Shmueli, E. (2023). A review on fairness in machine learning. *ACM Computing Surveys*, *55*(3), 1–44. https://doi.org/10.1145/3494672

Rafii, M. S., & Aisen, P. S. (2023). Detection and treatment of alzheimer's disease in its preclinical stage [Epub 2023 May 18]. *Nature Aging*, *3*(5), 520–531. https://doi.org/10.1038/s43587-023-00410-4

Ren, J., Xu, H., Liu, Y., Cui, Y., Wang, S., Yin, D., & Tang, J. (2024). A Robust Semantics-based Watermark for Large Language Model against Paraphrasing. *Findings of the Association for Computational Linguistics: NAACL 2024*, 613–625. https://doi.org/10.18653/v1/2024.findings-naacl.40

Santander-Cruz, Y., Salazar-Colores, S., Paredes-García, W. J., Guendulain-Arenas, H., & Tovar-Arriaga, S. (2022). Semantic Feature Extraction Using SBERT for Dementia Detection. *Brain Sciences*, *12*(2), 270. https://doi.org/10.3390/brainsci12020270

Searle, T. (2020). *Interspeech 2020 - alzheimer's dementia recognition through spontaneous speech challenge* [GitHub repository. Source code (Jupyter Notebook)]. https://github.com/tomolopolis/ADReSS_Challenge/blob/master/Analysis.ipynb

Shamei, A., Liu, Y., & Gick, B. (2023). Reduction of Vowel Space in Alzheimer's Disease. *JASA Express Letters*, *3*(3), 035202. https://doi.org/10.1121/10.0017438

Van Haastrecht, M., Sarhan, I., Yigit Ozkan, B., Brinkhuis, M., & Spruit, M. (2021). SYMBALS: A systematic review methodology blending active learning and snowballing. *Frontiers in Research Metrics and Analytics*, *6*, 685591. https://doi.org/10.3389/frma.2021.685591

Verma, M., & Howard, R. J. (2012). Semantic Memory and Language Dysfunction in Early Alzheimer's Disease: A Review. *International journal of geriatric psychiatry*, *27*(12), 1209–1217.

Wegner, P., Balabin, H., Ay, M. C., Bauermeister, S., Killin, L., Gallacher, J., Hofmann-Apitius, M., Salimi, Y., Initiative, A. D. N., Initiative, J. A. D. N., Study, A. B., Borders, A. R. W., & Consortium, E. (2024). Semantic Harmonization of Alzheimer's Disease Datasets Using AD-Mapper. *Journal of Alzheimer's Disease*, *99*(4), 1409–1423. https://doi.org/10.3233/JAD-240116

Wen, B., Wang, N., Subbalakshmi, K., & Chandramouli, R. (2023). Revealing the Roles of Part-of-Speech Taggers in Alzheimer Disease Detection: Scientific Discovery Using One-Intervention Causal Explanation. *JMIR Formative Research*, *7*, e36590. https://doi.org/10.2196/36590

Williams, E., McAuliffe, M., & Theys, C. (2021). Language changes in Alzheimer's disease: A systematic review of verb processing. *Brain and Language*, *223*, 105041. https://doi.org/10.1016/j.bandl.2021.105041

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., . . . Rush, A. M. (2020). HuggingFace's Transformers: State-of-the-art Natural Language Processing. https://arxiv.org/abs/1910.03771

Wu, J., Liu, Y., & Zong, C. (2024). F-MALLOC: Feed-forward Memory Allocation for Continual Learning in Neural Machine Translation. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 7180–7192. https://doi.org/10.18653/v1/2024.naacl-long.398

Zhang, W. (2024). Ai fairness in practice: Paradigm, challenges, and prospects. *AI Mag.*, *45*(3), 386–395. https://doi.org/10.1002/aaai.12189

Zhang, X., Sun, J., Hong, S., & Li, T. (2024). Amanda: Adaptively Modality-Balanced Domain Adaptation for Multimodal Emotion Recognition. *Findings of the Association for Computational Linguistics: ACL 2024*, 14448–14458. https://doi.org/10.18653/v1/2024.findings-acl.859

Zhu, Z., Novikova, J., & Rudzicz, F. (2019). Deconfounding age effects with fair representation learning when assessing dementia [Version 4]. *arXiv preprint arXiv:1807.07217v4*. https://arxiv.org/abs/1807.07217v4

Appendix A: Additional Explainability Plots for ADReSS

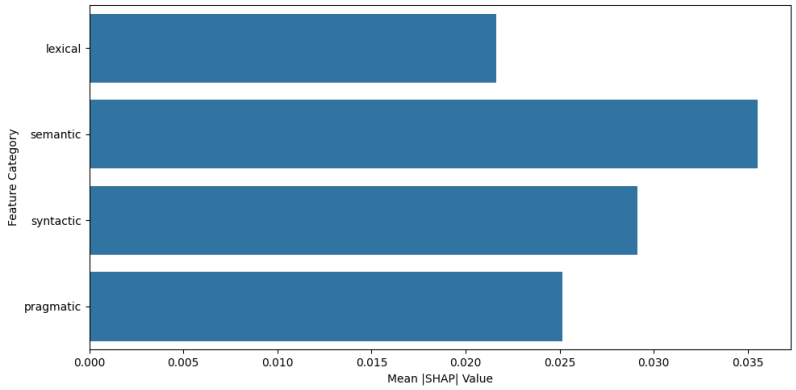
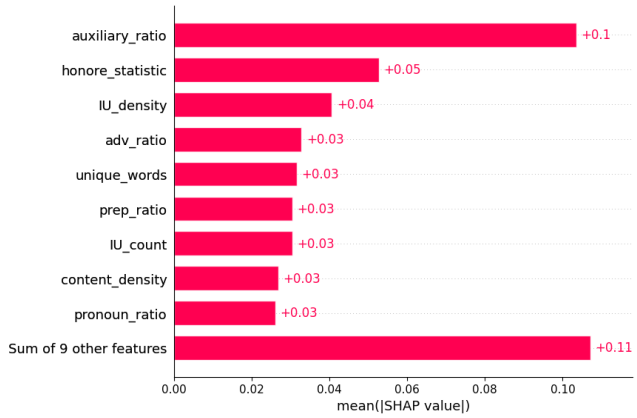
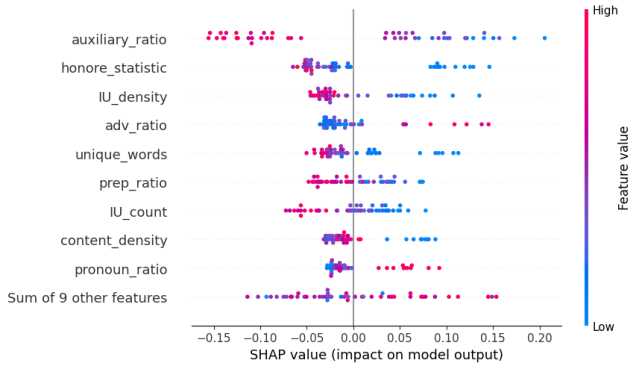


Figure 13: Average SHAP values grouped by linguistic feature category: lexical, semantic, syntactic, and pragmatic. Semantic and syntactic features (e.g., IU_count, mean_sentence_length, clauses_per_sentence) have the greatest overall impact on the model's predictions. This indicates that vocabulary richness and syntactic structure play key roles in distinguishing AD speech from control samples.



(a) Feature-level SHAP values showing the mean absolute contribution of each feature. Auxiliary_ratio and honore_statistic are among the most influential.



(b) Instance-level SHAP value distribution for the top features. High auxiliary_ratio (in red) reduces AD scores, being linked to control speech.

Figure 14: SHAP analysis for ADReSS dataset of both global feature importance and individual prediction impact for the AD classification RF model.

Appendix B: Explainability Results for the Pitt Dataset

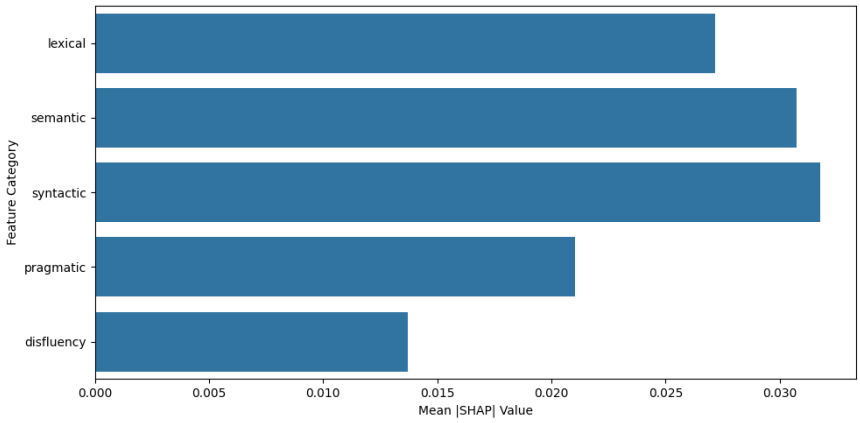
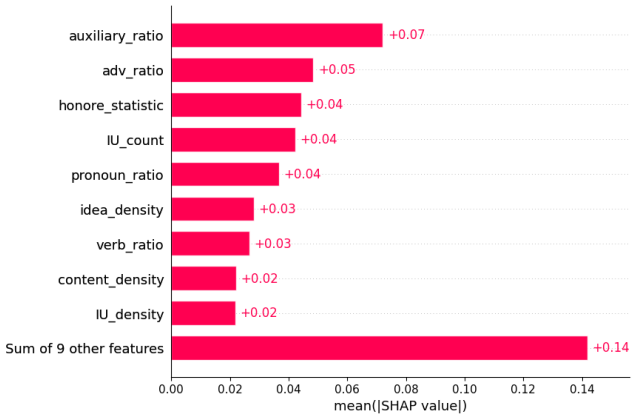
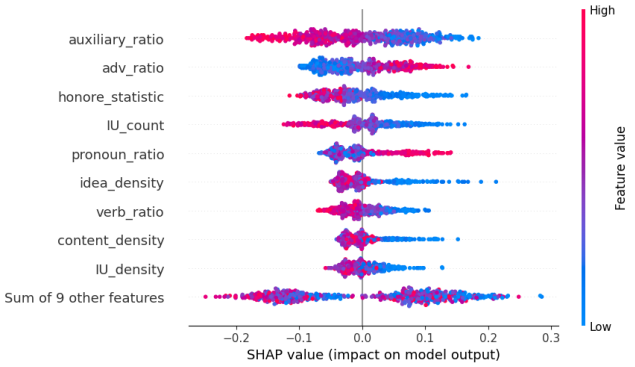


Figure 15: Average SHAP values grouped by linguistic feature category for the Pitt dataset. Syntactical features (e.g., parse_tree_depth, clauses_per_sentence) dominates model predictions. This suggests that sentence structure and grammatical information are key for detecting AD in this dataset. Semantic features are almost as relevant for Pitt dataset.



(a) Feature-level SHAP values for the Pitt model. Overall trends resemble those in ADReSS, except some features are more influential here, such as idea_density.



(b) Instance-level SHAP value distribution for the Pitt dataset. Features are ranked by importance, with SHAP values showing their direction and magnitude of impact on AD predictions. As with ADReSS, higher auxiliary_ratio lowers AD prediction.

Figure 16: SHAP analysis on the Pitt dataset of both global feature importance and individual prediction impact for the AD classification RF model.

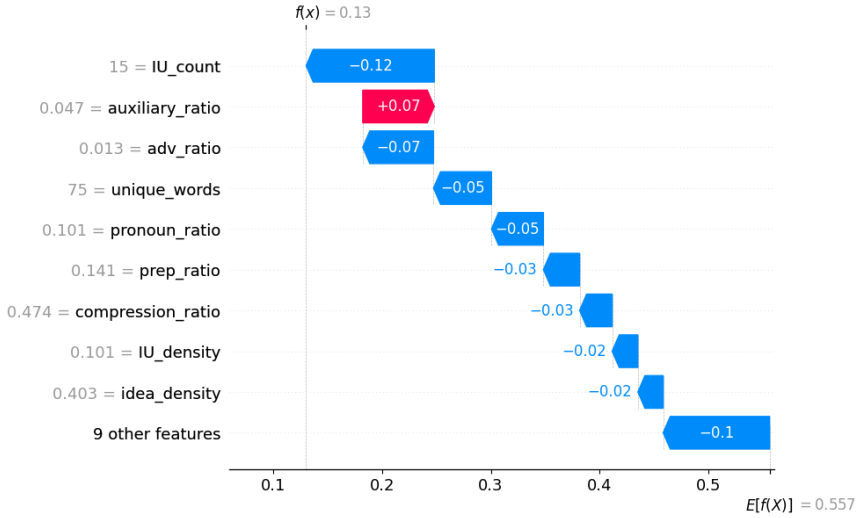


Figure 17: SHAP waterfall plot for an individual prediction from the RF model trained on the Pitt dataset. The plot begins with the model's expected output (0.557), and sequentially adds or subtracts SHAP values from features to reach the final predicted probability of 0.13 for AD class. Positive SHAP values (red) increase AD prediction while negative ones (blue) push it toward control class. Auxiliary_ratio is the most influential positive contributor, while features like pronoun_ratio and IU_count decrease AD prediction.

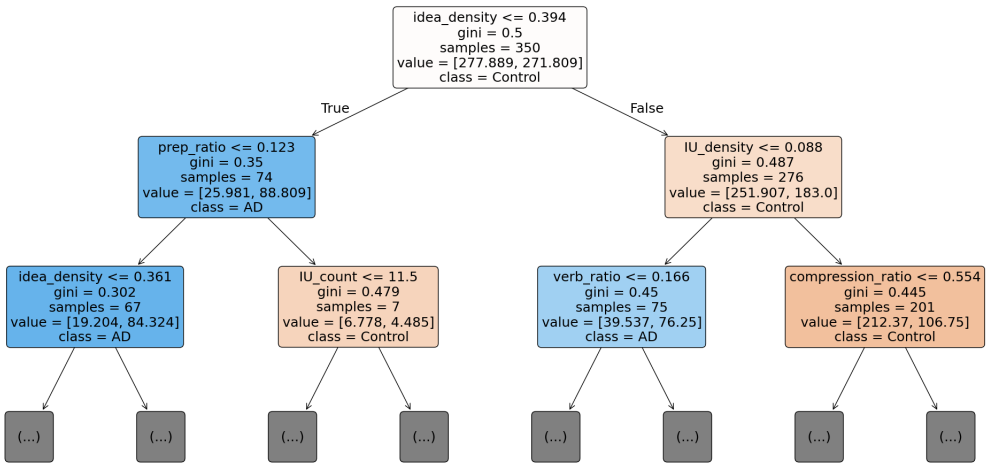


Figure 18: Visualization of one decision tree with depth 2 from the trained Random Forest model on the Pitt Corpus dataset. The tree shows key linguistic splits, such as idea_density, compression_ratio, and IU_count, are used to classify AD vs control speech samples.

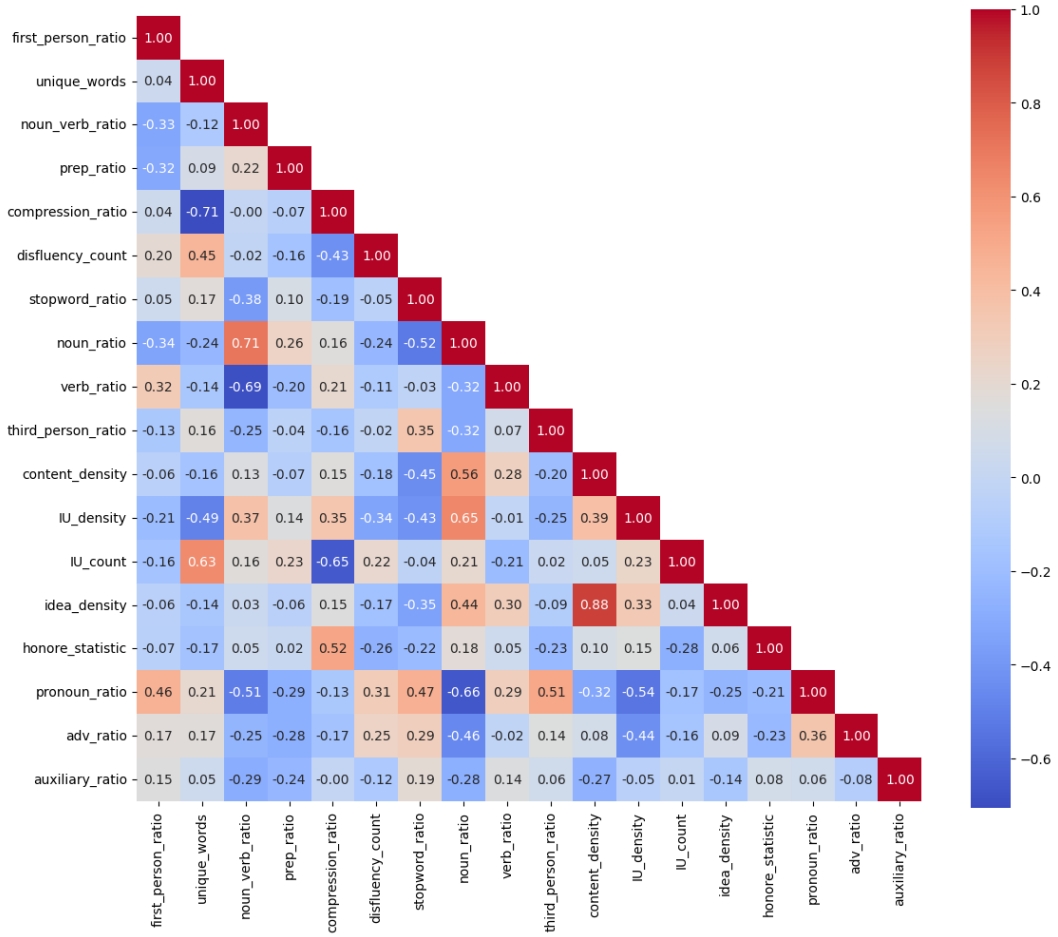


Figure 19: Pearson correlation heatmap of the top 18 features used in the RF model for the Pitt dataset. Each cell represents the correlation coefficient between a pair of features; red indicating a strong positive correlation and blue, strong negative correlation. Notable relationships include high positive correlations between IU_density and noun_ratio (0.65) and strong negative between unique_words and compression_ratio (-0.71).